University of Zurich[UZH]

Department of Informatics

Sensing and Indicating Interruptibility in Office Workplaces

Dissertation submitted to the Faculty of Business,
Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktorin der Wissenschaften, Dr. sc.
(corresponds to Doctor of Science, PhD)

presented by
Manuela Züger
from Schübelbach, SZ, Switzerland

approved in October 2018

at the request of
Prof. Thomas Fritz, Ph.D.
Prof. Joanna McGrenere, Ph.D.
Prof. Chat Wacharamanotham, Ph.D.

University of Zurich[UZH]

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, October 24, 2018

Chairman of the Doctoral Board: Prof. Dr. Sven Seuken

Acknowledgments

It is my pleasure to thank to everyone who made this thesis possible and supported me along this journey.

First and foremost, I would like to thank Thomas Fritz, my advisor, for giving me the opportunity to pursue my doctoral studies and guiding me during the past four years. I very much appreciate his challenging me, believing in me, and giving me highly valuable advice on many things.

I am also very grateful to Joanna McGrenere for taking on the role as external examiner, and for dedicating her time to giving insightful recommendations and helpful advice at various occasions.

I am especially grateful to Chat Wacharamanotham for being part of my PhD committee, and for spending his time organizing feedback groups and providing invaluable feedback during the paper writing process, despite his busy schedule. At the same time, I would like to thank Elaine Huang, Helen Ai He, Christian Remy and Chia-Kai Yang from the ZPAC group, for their highly valuable feedback before presentations and paper deadlines.

I would like to give special thanks to Harald Gall, for the many interesting discussions at various occasions and in particular for giving me the opportunity to be part of the supportive and fun SEAL group.

I appreciate that I had the opportunity to collaborate with great researchers from academia and industry during my PhD, including David Shepherd, Christopher Corley, Boyang Li, Vinay Augustine from, Patrick Francis, Nicholas Kraft, and Will Snipes from ABB Corporate Research, and André Meyer and Sebastian Müller from the University of Zürich.

This thesis would not have been possible without the 482 study participants. Thank you all for spending your time to collect the data which is the basis of this work.

Special thanks go to my lab mates, the members of the SEAL group: Carol Alexandru, Pooyan Behnamghader, Martin Brandtner, Jürgen Cito, Adelina Ciurumelea, Giovanni Grano, Christian Inzinger, Katja Kevic, Christoph Laaber, Philipp Leitner, André Meyer, Sebastian Müller, Sebastiano Panichella, Sebastian Proksch, Gerald Schermann, and Carmine Vassallo. Thank you for your support during all the ups and downs I experienced during my PhD, for refreshing walking meetings, UNO and push-up breaks, for the fun times and discussions we had in the office, after work, at conferences and at retreats, and for your highly valuable feedback on my work.

Last but not least I would like to thank my family, Vreni, Wisi, Roland and Isabel for always and unconditionally supporting and encouraging me; my boyfriend, Dominic, for the countless positive words, relaxing times in the nature, and constant moral support; and my dance family and friends, for listening to my numerous PhD stories, dancing with me to boost my creativity, and your endless contagious positive energy.

Abstract

In office workplaces, interruptions by co-workers, emails or instant messages are common. Many of these interruptions are useful as they might help resolve questions quickly and increase the productivity of the team. However, knowledge workers interrupted at inopportune moments experience longer task resumption times, lower overall performance, more negative emotions, and make more errors than if they were to be interrupted at more appropriate moments.

To reduce the cost of interruptions, several approaches have been suggested, ranging from simply closing office doors to automatically measuring and indicating a knowledge worker's interruptibility—the availability for interruptions—to co-workers. When it comes to computer-based interruptions, such as emails and instant messages, several studies have shown that they can be deferred to automatically detected breakpoints during task execution, which reduces their interruption cost. For in-person interruptions, one of the most disruptive and time-consuming types of interruptions in office workplaces, the predominant approaches are still manual strategies to physically indicate interruptibility, such as wearing headphones or using manual busy lights. However, manual approaches are cumbersome to maintain and thus are not updated regularly, which reduces their usefulness.

To automate the measurement and indication of interruptibility, researchers have looked at a variety of data that can be leveraged, ranging from contextual data, such as audio and video streams, keyboard and mouse interaction data, or task characteristics all the way to biometric data, such as heart rate data or eye traces. While studies have shown promise for the use of such sensors, they were

predominantly conducted on small and controlled tasks over short periods of time and mostly limited to either contextual or biometric sensors. Little is known about their accuracy and applicability for long-term usage in the field, in particular in office workplaces. In this work, we developed an approach to automatically measure interruptibility in office workplaces, using computer interaction sensors, which is one type of contextual sensors, and biometric sensors. In particular, we conducted one lab and two field studies with a total of 33 software developers. Using the collected computer interaction and biometric data, we used machine learning to train interruptibility models. Overall, the results of our studies show that we can automatically predict interruptibility with high accuracy of 75.3%, improving on a baseline majority classifier by 26.6%.

An automatic measure of interruptibility can consequently be used to indicate the status to others, allowing them to make a well-informed decision on when to interrupt. While there are some automatic approaches to indicate interruptibility on a computer in the form of contact list applications, they do not help to reduce in-person interruptions. Only very few researchers combined the benefits of an automatic measurement with a physical indicator, but their effect in office workplaces over longer periods of time is unknown. In our research, we developed the FlowLight, an automatic interruptibility indicator in the form of a traffic-light like LED placed on a knowledge worker's desk. We evaluated the FlowLight in a large-scale field study with 449 participants from 12 countries. The evaluation revealed that after the introduction of the FlowLight, the number of in-person interruptions decreased by 46% (based on 36 interruption logs), the awareness on the potential harm of interruptions was elevated and participants felt more productive (based on 183 survey responses and 23 interview transcripts), and 86% remained active users even after the two-month study period ended (based on 449 online usage logs).

Overall, our research shows that we can successfully reduce in-person interruption cost in office workplaces by sensing and indicating interruptibility. In addition, our research can be extended and opens up new opportunities to further support interruption management, for example, by the integration of other more

accurate biometric sensors to improve the interruptibility model, or the use of the model to reduce self-interruptions.

Zusammenfassung

Unterbrechungen von Mitarbeitern, E-Mails oder Chatnachrichten sind an heutigen Büroarbeitsplätzen alltäglich. Viele dieser Unterbrechungen sind hilfreich und erhöhen die Produktivität des Teams, da so manche Probleme schneller gelöst werden können. Jedoch können Unterbrechungen auch zu ungünstigen Zeitpunkten auftreten, wie etwa wenn eine Person sehr fokussiert ist. Dies resultiert in einer geringeren Arbeitsleistung, langen Wiedereinarbeitungszeiten, negativen Emotionen und mehr Fehlern, was in Summe hohe Kosten verursachen kann.

Es wurden verschiedene Ansätze entwickelt, um die hohen Kosten von Unterbrechungen zu reduzieren. Diese spannen den Bogen von einfachen Strategien wie die Bürotüre zu schliessen, bis zur automatischen Messung und Anzeige der Unterbrechbarkeit—der Verfügbarkeit für Unterbrechungen. Für Unterbrechungen am Computer haben Forscher bereits gezeigt, dass diese als weniger störend empfunden werden, wenn sie automatisch erst am Ende der Aufgabe angezeigt werden. Für persönliche Unterbrechungen, welche zu den störendsten und zeitintensivsten Arten von Unterbrechungen im Büro gehören, sind die vorherrschenden Optimierungsstrategien manuelle Ansätze, wie beispielsweise Kopfhörer aufzusetzen oder ein von Hand gesteuertes Ampellicht. Solche manuellen Ansätze sind jedoch wartungsaufwändig und werden nur selten aktualisiert, was sie weniger nützlich macht.

Um die aktuelle Unterbrechbarkeit automatisch messen und anzeigen zu können, haben Forscher bisher verschiedene Datenquellen untersucht. Dazu gehören hauptsächlich kontextuelle Sensordaten wie Audio- und Video-Streams, Tastatur-

und Mausinteraktionsdaten oder Charakteristika der aktuellen Aufgabe. Einige wenige Studien haben ausserdem biometrische Sensoren wie Pulsmonitoren oder Eye Tracker verwendet. Die existierenden Studien haben gezeigt, dass diese Sensoren Potential für die Messung von Unterbrechbarkeit haben. Jedoch wurden die zumeist kurzen Studien oft in einer kontrollierten Umgebung durchgeführt, und fokussierten sich entweder auf kontextuelle oder biometrische Sensoren. Somit ist nur wenig über die Anwendbarkeit und Genauigkeit dieser Sensoren bei längerfristigem Einsatz an Büroarbeitsplätzen bekannt. In dieser Arbeit entwickelten wir eine Methode, um Unterbrechbarkeit an Büroarbeitsplätzen automatisch zu messen und verwendeten dafür eine Kombination aus Computerinteraktionsdaten, welche eine Art von kontextuellen Sensordaten darstellen, und biometrischen Daten. Um diese Daten zu gewinnen, haben wir eine Labor- und zwei Feldstudien mit insgesamt 33 Softwareentwicklern durchgeführt. Mit den gesammelten Computerinteraktions- und biometrischen Daten trainierten wir Machine Learning Modelle für Unterbrechbarkeit. Die Resultate unserer Studien zeigen, dass wir Unterbrechbarkeit automatisch und mit einer hohen Genauigkeit von 75.3% vorhersagen können, was 26.6% besser als der Referenzwert eines Majority Classifiers ist.

Ein automatisches Mass für Unterbrechbarkeit kann dann anderen Mitarbeitern angezeigt werden, wodurch diese eine besser informierte Entscheidung treffen können, wann Sie ihre Arbeitskollegen unterbrechen können. Existierende Ansätze zur automatischen Anzeige der Unterbrechbarkeit am Computer helfen nicht, um persönliche Unterbrechungen zu reduzieren. Nur sehr wenige Forscher haben den Vorteil einer automatischen Messung mit einer physischen Anzeige kombiniert, jedoch ist deren längerfristiger Einfluss an Büroarbeitsplätzen nicht bekannt. In unserem Ansatz zeigen wir die Unterbrechbarkeit mit dem FlowLight an, einer automatischen Unterbrechbarkeits-Anzeige in Form einer physischen Ampel-ähnlichen Lampe, welche am Arbeitsplatz befestigt wird. Wir haben das FlowLight in einer grossen Feldstudie mit 449 Teilnehmern aus 12 Ländern evaluiert. Die Evaluation hat ergeben, dass die Anzahl der persönlichen Unterbrechungen nach der Einführung von FlowLight um 46% sank (basierend auf 36 Unterbrechungsaufzeichnungen). Ausserdem wurde den Teilnehmern die

potentiellen Unterbrechungskosten zunehmend bewusst und sie fühlten sich produktiver (basierend auf 183 Umfrageantworten und 23 Interviewtranskripten) und ein Grossteil von ihnen (86%) nutzen das FlowLight nach dem Ende der zweimonatigen Studie weiter (basierend auf 449 Online Nutzungs-Aufzeichnungen).

Zusammenfassend zeigen unsere Ergebnisse, dass wir durch die automatische Messung und Anzeige der individuellen Unterbrechbarkeit die Kosten von persönlichen Unterbrechungen an Büroarbeitsplätzen erfolgreich verringern können. Zusätzlich kann unsere Forschung erweitert werden und eröffnet neue Möglichkeiten, um den Umgang mit Unterbrechungen noch besser zu unterstützen, beispielsweise durch die Integration weiterer oder genauerer biometrischen Sensoren, oder durch die Nutzung des Modells zur Reduktion von Selbst-Unterbrechungen.

Contents

List of Figures

List of Tables

Acronyms

AUC area under curve

API application programming interface

BVP blood volume pulse

CV cross-validation

DnD do not disturb

EDA electrodermal activity

ECG electrocardiogram or electrocardiography

EEG electroencephalogram or electroencephalography

EMG electromyogram or electromyography

h hour(s)

HR heart rate

HRV heart rate variability

Hz Hertz

IBI interbeat interval

IDE integrated development environment

IM instant messaging

MI mutual information

min minute(s) or minimum

max maximum

PNN20 percentage of successive IBIs with a difference greater than 20ms

PNN20 percentage of successive IBIs with a difference greater than 50ms

PPG photoplethysmography

RMSSD root mean square of successive IBI differences

RQ research question

s second(s)

SCL skin conductance level

SDNN standard deviation of normal-to-normal heartbeat intervals

Stdev or std. dev. standard deviation

1

Synopsis

In today's collaborative work environments, knowledge workers are constantly facing interruptions, such as instant message alerts, emails or co-workers asking a question in person [González and Mark, 2004, Chong and Siino, 2006, Iqbal and Horvitz, 2007]. Many of these interruptions are necessary to share knowledge and resolve problems quickly [Isaacs et al., 1997]. Yet, the timing of the interruption can have a big impact on its disruptiveness [Adamczyk and Bailey, 2004, Bailey and Konstan, 2006]. Several studies have demonstrated the negative effects of interruptions, ranging from higher error rates and lower overall performance to increased stress and frustration; especially when the interruptions happen at inopportune moments such as during highly focused work [Bailey et al., 2001, Czerwinski et al., 2000, Mark et al., 2008]. Due to the negative effects and high cost of interruptions, researchers have developed approaches to postpone interruptions at inopportune moments to more suitable times. It has been shown that computer-based interruptions such as emails and instant messages can be postponed to automatically detected task switches during computer

work, which are moments of low cognitive load and thus more suitable for interruptions [Bailey and Iqbal, 2008, Iqbal and Bailey, 2008]. For in-person interruptions, which is one of the most disruptive and time-consuming type of interruptions in office workplaces [Sykes, 2011, González and Mark, 2004], a possible approach is to measure a person's interruptibility—the availability for interruptions—continuously, and indicate this state to potentially interrupting co-workers [Begole et al., 2004, Bjelica et al., 2011]. Such an automatic indicator can enable well-informed decisions about when and how to interrupt a co-worker and potentially decreases the number of interruptions that occur at inopportune moments. While researchers have started to explore approaches to automatically sense and indicate interruptibility, little is known about the feasibility, accuracy and best sensing techniques of such a measurement in office workplaces, and whether it can be used to successfully reduce in-person interruption cost.

Previous research has examined the use of various sensors to measure a person's interruptibility. These sensors can predominantly be categorized as either contextual or biometric sensors. Prior work on sensing interruptibility mostly examined the use of contextual data that spans from audio and video streams over keyboard and mouse actions, active window information to task characteristics (e.g. [Hudson et al., 2003, Fogarty et al., 2005b, Fogarty et al., 2005a, Iqbal and Bailey, 2006]). In contrast to sensing contextual data, biometric sensors are more physically invasive but have the advantage of providing more flexibility without being bound to a specific task, computer or location; especially now that biometric sensors have become more accessible, more accurate and less invasive. Researchers have conducted lab studies using biometric sensors such as electrodermal activity (EDA), heart rate (HR) or electroencephalography (EEG) sensors to measure cognitive load and emotional aspects (e.g., [Nourbakhsh et al., 2012, Grimes et al., 2008]). Under the assumption that moments of high cognitive load or stress correlate with low interruptibility, a few studies have started to examine the use of biometric sensors to measure interruptibility in lab settings [Chen et al., 2007, Bailey and Iqbal, 2008]. Overall, prior work on sensing interruptibility has predominantly focused on short controlled lab experiments and on either contextual or biometric sensors. Little is known

about the feasibility of a continuous and automatic interruptibility measurement for knowledge workers working on their own tasks and in their usual work environment. In our research, we aim at determining a highly accurate approach to measure interruptibility in office workplaces and we therefore examine a broad range of computer interaction sensors (as one type of contextual sensors) and biometric sensors.

An automatic interruptibility measurement can be indicated to co-workers with the goal of reducing the cost of in-person interruptions by postponing them to more opportune moments. To better manage interruptions, researchers and practitioners developed interruptibility indicators in the form of computer-based applications such as contact lists along with interruptibility information (e.g. [Tang et al., 2001, Begole et al., 2004, Fogarty et al., 2004]), or in the form of physical indicators such as closed office doors or manual busy lights [Sykes, 2011, Embrava, 2016]. Evaluations of computer-based indicators did not reveal any changes to the cost of in-person interruptions, probably since the contact-list style applications can easily be hidden behind other applications and thus forgotten at communication initiation [Begole et al., 2004, Hincapié-Ramos et al., 2011a]. Physical indicators have the advantage of being more prominently visible. However, existing approaches such as manual busy lights rely on manual status updates and thus are generally too cumbersome to maintain [Milewski and Smith, 2000]. Very few approaches have looked at combining physical interruptibility indicators with automatic interruptibility measures to reduce the cost of in-person interruptions [Hincapié-Ramos et al., 2011b, Bjelica et al., 2011] and there is no knowledge on the long-term effects of such approaches. In our research, we focus on building an automatic interruptibility indicator in the form of a physical traffic-light like LED, placed directly at the desk of each knowledge worker. We then evaluated the light's effect on the cost of interruptions.

To summarize, the goal of our research is to reduce the cost of interruptions in office workplaces with a specific focus on in-person interruptions. To achieve this goal, we developed an accurate and minimally invasive approach to measure interruptibility and to provide awareness on interruptibility to co-workers.

This leads us to the following ***hypotheses***:

H1: A combination of computer interaction and biometric sensors can be used to measure knowledge workers' interruptibility in office workplaces automatically with high accuracy.

H2: An automatic interruptibility indicator in the form of a physical light can reduce the cost of in-person interruptions in an office work environment.

To investigate these hypotheses, we conducted one lab and two field studies to evaluate *Hypothesis H1* and one field study to evaluate *Hypothesis H2*. We focused on two main research questions that are described in more detail in the next section (Section 1.1). Section 1.2 presents the technical approach we used to answer our research questions. The findings of our research are presented in section 1.3, followed by the threats to validity (Section 1.4). Then we discuss challenges (Section 1.5), opportunities and potential future work (Section 1.6), discuss related work (Section 1.7), summarize our work and contributions (Section 1.8) and describe the roadmap of this thesis (Section 1.9).

1.1 Research Questions

To validate our hypotheses, we will examine the following ***research questions*** on sensing and indicating interruptibility:

RQ 1: Can we measure the interruptibility of knowledge workers automatically with high accuracy?

RQ 1a: Can we use biometric sensors to measure interruptibility in the lab and field automatically with high accuracy?

RQ 1b: Which combination of computer interaction and biometric sensors is best to measure interruptibility in office workplaces with high accuracy?

RQ 2: Can we reduce in-person interruption cost with a physical and automatic interruptibility indicator?

To answer these questions, we conducted a range of lab and field studies. Table 1.1 depicts an overview of the studies along with the chapter describing the corresponding research in detail.

Our aim for *RQ 1* is to develop an approach to measure interruptibility in office workplaces continuously and accurately. Since biometric sensors have previously shown great potential in measuring cognitive and emotional states, we start with exploring the feasibility of using biometric sensors to measure interruptibility (*RQ 1a*). We extend existing work from lab studies to the field, in particular by evaluating a biometric interruptibility measure in office workplaces. First, we conducted a one-hour lab study with 10 graduate students working on predefined software development tasks, followed by a two-hour field study with 10 professional software developers working on their own tasks. With *RQ 1b*, we build upon and extend our results from *RQ 1a* and extend the biometric interruptibility measurement with one type of contextual sensors, namely computer interaction sensors, since these can be automated and are little invasive to use on a daily basis. With data collected in the field from 13 professional software developers over two weeks, we investigated each sensor's accuracy and the overall generalizability of the interruptibility measurement.

To answer *RQ 2*, we developed an approach that combines an interruptibility measure based on computer interaction sensors with a physical interruptibility indicator light. In a large-scale field study with 449 participants including software developers, project managers and other job roles from 15 sites of one multi-national company located in 12 countries, we investigated whether such an interruptibility indicator light can decrease the number of (disruptive) in-person interruptions. We further investigated whether such an indicator changes the behavior or interactions of knowledge workers and whether it has the potential to be adopted in the long term.

RQ	Study	Chapter	Type	Length	Participants	Sensors
1a	1	2	Lab	1 hour	10 graduate students	Empatica E3, Neurosky Mindband
	2		Field	2 hours	10 software developers	Empatica E3, Neurosky Mindband
1b	3	3	Field	2 weeks	13 software developers	Computer monitoring, Polar H7, Fitbit Charge 2
2	4	4	Field	2 months	449 knowledge workers	Computer monitoring

Table 1.1: Studies conducted for this thesis along with the corresponding research question (RQ), chapter, whether it was conducted in the lab or field, the duration of the study, the participants and the sensors used to determine interruptibility.

1.2 Research Approach and Study Setup

In the following, we present our approach to sense (*RQ 1*) and indicate (*RQ 2*) interruptibility to reduce interruption cost in office workplaces.

1.2.1 RQ 1: Sensing Interruptibility

To build an automatic and real-time interruptibility measurement, we conducted one lab and two field studies (see Table 1.1). Over the course of the studies, we altered the number and variety of sensors used, increased the duration of the data collection, and moved from controlled lab studies to field studies in office workplaces. All of our studies follow a similar procedure. In the following, we outline our data collection and cleaning, feature extraction, normalization, and machine learning processes used to predict interruptibility at any given moment in time. Our data collection[1] and analysis[2] software for *study 3* are available online.

[1] `https://pluto.ifi.uzh.ch/PersonalAnalytics/`
[2] `https://zenodo.org/record/1118966`

Study Setup. For our studies, participants were asked to either work on predefined coding tasks (lab study) or on their own tasks (field studies) and were prompted at random intervals to rate their interruptibility. For *study 1* and *study 2*, these prompts were displayed on a tablet computer. For *study 3*, they were displayed at the computer where the participant was working on. For *study 1* and *study 2*, we focused on the use of biometric sensors to predict interruptibility. For *study 3*, we added computer interaction sensors and increased the data collection period to two weeks.

In our research, we used a variety of biometric sensors, including a Neurosky[3] MindBand to collect EEG and eye blink data; an Empatica[4] E3 to collect EDA, skin temperature, and blood volume pulse (BVP) data; a Fitbit[5] Charge 2 to collect HR data with an optical sensor, as well as movement and sleep data; and a Polar[6] H7 to measure HR data with an electrocardiography (ECG)-based sensor. To gather computer interaction data, we used a monitoring tool developed by Meyer et al. called WorkAnalytics [Meyer et al., 2017b].

For all our studies, we further conducted short interviews to learn more about the participants' experience with interruptions and the sensors, and collected demographic data.

Data Cleaning. After collecting the data, we needed to anonymize and clean it. To anonymize the data, we redacted any potentially identifying texts occurring in the computer interaction data by replacing them with a placeholder, e.g. replacing an email address with "<email2>". Further, we needed to remove noise from the collected raw data (in particular the biometric data). As an example, we applied a 50 Hz notch filter to remove signal noise caused by overhead lights from the raw EEG sensor data.

Feature Extraction. The collected raw data is often not meaningful by itself and we need to extract features from it. These features can then be used with

[3] `http://neurosky.com`
[4] `https://www.empatica.com`
[5] `https://www.fitbit.com`
[6] `https://www.polar.com`

a machine learning approach to predict interruptibility. We chose our features based on literature linking the feature to interruptibility or other related states such as a high cognitive load or stress level. All extracted feature groups and examples of some of the features are presented in Table 1.2.

To calculate a feature, such as the mean heart rate variability (HRV), we first needed to determine and segment the continuous data into time windows. Since there is no common standard of time window length used per feature, we first determined the optimal time window duration per feature. Using the optimal time window duration for segmenting the data stream, we then extracted features from our wide variety of cleaned data.

As an example for a feature extracted from computer interaction data, we semi-automatically coded the active application on the computer into predefined application categories, such as email or planning. Then we calculated the time spent in each category during the time window, since such features have already been linked to productivity and interruptibility by prior work [Meyer et al., 2014, Iqbal and Bailey, 2007]. As an example for a biometric feature, we split the EDA signal into its phasic and tonic components, and calculated the mean and peak related metrics. These components of the EDA data have previously been linked to arousal and specific emotions [Boucsein, 2012].

Normalization. As computer interaction and biometric data is usually substantially different across different individuals, we needed to normalize the data to build a model across multiple participants. For *study 1* and *study 2*, we collected individual baseline measures for participants while they watched a calming fish tank video for two minutes, and used these measures to normalize the feature values. For the long study (*study 3*), we normalized the features using standardization to center and scale the features to unit variance, resulting in a distribution with a mean of 0 and a standard deviation of 1.

Biometric Feature Categories and Samples	
Brain	EEG Frequency bands (e.g. α or β) and combinations (e.g. α/γ), Neurosky's Attention and Mediation scores (e.g. max. value in time window) [Kramer, 1991, Lee and Tan, 2006]
Eye	Eye blinks (e.g. # per minute) [Manoilov, 2007]
Heart	BVP Amplitude (e.g. max peak amplitude), HR (e.g. mean), HRV (e.g. PNN50) [Peper et al., 2007, Camm et al., 1996]
Skin	EDA Phasic signal (e.g. mean peak amplitude), EDA tonic signal (e.g. mean), temperature (e.g. mean) [Boucsein, 2012, Nourbakhsh et al., 2012]
Movement	Steps (e.g. number per minute) [Ho and Intille, 2005, Fisher and Simmons, 2011]
Sleep	Duration (e.g. total minutes), quality (e.g. restless minutes) [Pilcher et al., 1997, Rosekind et al., 2010]
Computer Interaction Feature Categories and Samples	
Time	Current time (e.g. hour of day), circadian rhythm (e.g. hour arrived at work) [Mark et al., 2014, Visuri et al., 2017]
Calendar	Meetings (e.g. # upcoming meetings) [Stern et al., 2011, Horvitz et al., 2002]
User Input	Keystrokes (e.g. # delete key presses), mouse clicks (e.g. # left clicks), mouse moves (e.g. moved pixels per minute), mouse scrolls (e.g. time spent scrolling) [Shrot et al., 2014, Kapoor and Horvitz, 2008]
Applications	Activity categories (e.g. time spent coding), focus duration (e.g. max. time in one window), activity switches (e.g. # window switches per minute) [Mirza et al., 2011, Iqbal and Bailey, 2007]

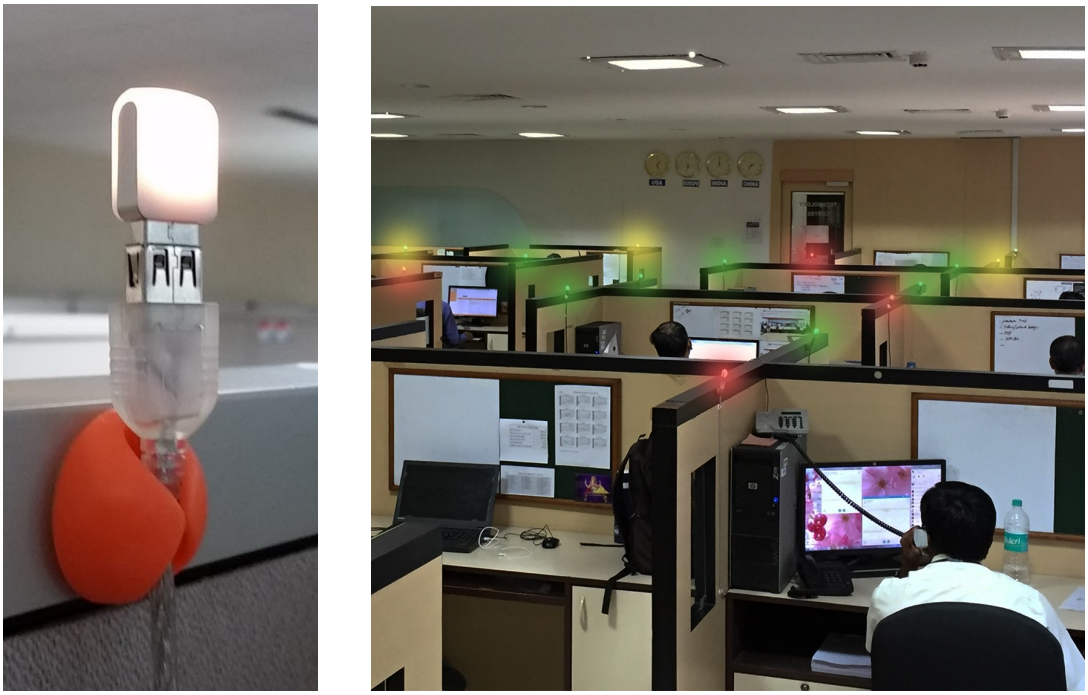
Table 1.2: Feature categories of the interruptibility model along with sample features (in brackets) and references to prior work using or defining these features.

Interruptibility Prediction. The last step was to train a machine learning classifier, and to use the resulting model to predict the participants' interruptibility ratings. We trained various classification algorithms and chose those with the highest accuracy: a Naïve Bayes classifier for *study 1* and *study 2* and a Random Forest classifier for *study 3*. Since we were interested in the effects of individual differences on the model, especially in the biometric data, we trained and examined both individual models (using data from only one participant) and general models (using data from all participants except the one used to validate the model).

1.2.2 RQ 2: Indicating Interruptibility

To reduce the number of in-person interruptions at inopportune moments, we developed the FlowLight, an application and a physical traffic light-like LED lamp that indicates interruptibility to co-workers in the office. The FlowLight also updates the computer-based instant messaging status, however, with its physical indicator it primarily focuses on in-person interruptions. In the following, we present the functionality and development of the FlowLight in more detail, followed by the study design to evaluate its effects.

FlowLight Approach. FlowLight consists of a computer application to automatically determine a user's interruptibility state and a physical LED light to indicate this state to co-workers. The physical LED light is mounted at the desk, cubicle wall or office door of a knowledge worker (see Figure 1.1). Similar to a traffic light, the light shows different colors to indicate a person's interruptibility: *available* as green, *busy* as red, *do not disturb (DnD)* as pulsating red, *idle or away* as yellow. The FlowLight application calculates the user's interruptibility state on the fly based on mouse and keyboard activity. We chose these data sources because they can be measured noninvasively and with limited privacy concerns. Specifically, the application determines a personalized measurement of interruptibility for each user based on heuristics for each type of input, the user's historical interaction data and a smoothing function. Based on insights from early pilot studies, the application sets the light to red for approximately



(a) FlowLight (b) Office setup with FlowLights

Figure 1.1: FlowLights mounted on knowledge workers' cubicle walls, photographed during the evaluation.

9% of the time spent on the computer (4% for pulsating red). Whenever the user's interruptibility status changes, the FlowLight application updates the color of the LED light as well as the Skype presence status. Additionally to the automatic status updates, the users had the possibility to manually change their status, e.g. when they wanted to focus on a urgent task.

Evaluation. To evaluate the FlowLight's ability to reduce in-person interruption cost, we conducted a large-scale field study with 449 participants of one multi-national company working in 12 different countries. In the study, we asked participants to self-report in-person interruptions for one week before we installed the FlowLight, and again after they familiarized themselves with the new system for a week. We further conducted surveys and interviews to learn about experienced costs and benefits of using the FlowLight. We were able to

collect a rich qualitative and quantitative dataset of 183 survey responses, 23 interview transcripts, 36 interruption logs, 47 FlowLight usage data logs and activity logs from all 449 participants.

1.3 Findings

In the following, we present our main research findings on sensing and indicating interruptibility. More details and further research insights are described in Chapters 2 to 5.

1.3.1 RQ 1: Sensing Interruptibility

Towards building an optimal interruptibility model for office workplaces, we evaluated the use of various computer interaction and biometric sensors, and evaluated the model's accuracy and applicability in the field.

Evaluating our interruptibility model trained with biometric data (*RQ 1a*), we found that the classifiers achieved an accuracy of 91.5% (two states) and 43.9% (five states) to predict interruptibility in the lab (*study 1*). Both classifiers also achieve a statistically significant improvement of 9.8% and 12.9% over a majority classifier. A majority classifier always predicts the most common class and is a commonly used as a baseline. In *study 2*, which was conducted in the field, only the classification into two states improved statistically significant over a majority classifier by 11.2%, achieving an accuracy of 78.6%. These results demonstrate the feasibility of using biometric sensors to predict interruptibility into two states both in the lab and field. The lower accuracy in the field compared to the lab study are not surprising given the sensitivity of the biometric sensors to external influences such as user movement and electrical interference from fluorescent lights.

To examine the use of various sensors over a longer time in the field, we used the larger data set collected in *study 3* over a two-week period with 13 professional software developers. The individually trained models achieve an overall accuracy of 75.3%, which is a 26.6% improvement over a majority classifier. A general

model—trained over all but one participant and evaluated on the remaining one—achieved an accuracy of 69.8%, improving over a majority classifier by 18%. This improvement shows that we can build and apply a general model for interruptibility with reasonable accuracy, which solves the cold-start-problem successfully. When comparing features, we found that the computer interaction features provided more predictive value compared to data from the Fitbit and Polar sensors (74.8% accuracy vs. 68.3%), and that a combination of all works best (75.3% accuracy). It must be taken into consideration that these findings are limited to times spent at or near the computer, since the interruptibility prompts were displayed at the computer and thus no ratings were collected while the participant was away from the computer. Our results suggest that computer interaction data can serve as a good starting point to sense interruptibility, but especially during tasks without extensive usage of the computer (e.g. reading or thinking), biometric sensors can complement computer interaction data well.

As an additional finding, we identified optimal time windows for the various features that we extracted from the raw data. We found that the optimal time window duration varies widely per feature. A short window of 10 seconds provided the most predictive power for biometric data in our shorter studies (*study 1* and *study 2*). In the two-week study (*study 3*), we also considered longer time windows up to 3 hours for the variety of features, and found that the optimal time window varies widely per feature, e.g. shorter time windows of 20 seconds for HR features, several minutes for application activity and up to 3 hours for calendar features.

To summarize, our findings on sensing interruptibility show that we can predict interruptibility with both computer interaction and biometric sensors with high accuracy and, in most cases, with statistically significant improvement over baseline. We further found that the optimal time window duration varies per feature, and that our interruptibility measurement is generalizable across individuals. Overall, these findings support *Hypothesis H1* and answer *RQ 1*.

1.3.2 RQ 2: Indicating Interruptibility

To reduce in-person interruption cost in office workplaces, we indicated interruptibility continuously to co-workers with the FlowLight and evaluated its effects on its users and their interruptions (*RQ 2*).

Our analysis of the 36 interruption logs showed that the FlowLight statistically significantly reduced the amount of in-person interruptions by 46%. A further major benefit of the FlowLight, as stated by participants in the survey (n=183) and interview (n=23), was that the small lights increased the awareness of interruption cost. For instance, one participant mentioned that *"the pilot increased the sensitivity to interruption. Team members think more about whether an interrupt is necessary and try to find a suitable time" (S45).* Furthermore, participants stated that the FlowLight increased their productivity: first, because it reduced the number of expensive interruptions and thus increasing their time available to work on their tasks; and second, because it encouraged them to focus or stay focused on their task either because they realized that their light was green for a while or because it just turned red. Overall, 85.5% of all users (n=449) also continued using the FlowLight even after the two-month study period ended.

In terms of the accuracy of the interruptibility status, 71% of the 183 survey respondents perceived the calculated status to be accurate. However, there is potential for improvement, especially in situations in which the participants' focus was high but they did not interact with the computer, e.g. when reading or sketching on a piece of paper. In these cases, biometric sensors worn by the user have the potential to increase the accuracy of the system, which is feasible as our findings on sensing interruptibility indicate.

Overall, the FlowLight's success and prolonged usage demonstrate that the combination of a physical indicator with an automatic interruptibility measurement is an effective means to reduce in-person interruption cost and to increase the awareness on the potential harm of interruptions as well as perceived productivity. These findings support *Hypothesis H2* and answer *RQ 2*.

1.4 Threats to Validity

In the following, we point out threats to the external, internal and construct validity of our research, and how we addressed and mitigated these threats.

External Validity. For our research addressing *RQ 1*, we chose software developers as one homogeneous group of knowledge workers to reduce confounding factors stemming from the nature of the work. Thus, our results may not generalize to different populations and environments. To mitigate this risk, we conducted multiple lab and field studies, had participants with various backgrounds from multiple different international companies, and had them work on defined tasks as well as on their own tasks for longer periods. While we believe that our work can be applied to other knowledge workers as well, more research is necessary to confirm this.

The generalizability of our evaluation of the FlowLight might be threatened by only having participants from one multi-national company and the limited study period of two months. We tried to mitigate this risk by having a high number of participants that further came from a large number of different sites and countries of the same multi-national corporation and from various business fields, ranging from software developers and other engineers to project managers. While we collected the interruption logs over a relatively short period of a few weeks, we collected other data such as interview transcripts, survey responses and usage logs after participants were using the FlowLights for a prolonged period of time up to several months. However, future research might look into effects of sustained usage of such systems over longer time spans.

Internal Validity. Biometric sensors bear a certain risk that they might record noisy and incorrect data, resulting in invalid results. This risk was higher in the field studies compared to the lab study, since external factors such as lighting conditions, environment temperature or the fit of the sensor can be less controlled. We mitigated this risk by using well-established sensors and applying noise cleaning and normalization techniques.

One possible threat to the internal validity of the FlowLight evaluation is that participants might have felt observed, as their FlowLight discloses their interruptibility to co-workers [McCarney et al., 2007]. They might feel inclined towards setting their FlowLight manually to the *busy* or *do not disturb* status frequently, which might reduce the amount of interruptions more than intended. We tried to mitigate this risk by instructing participants that it is important to be in the *available* status for a substantial time, as this allows their colleagues to approach them, e.g. when questions arise. Further, the FlowLight might actually increase the amount of interruptions in the beginning instead of reducing them, as the novel tool might provoke many questions caused by curiosity. We tried to mitigate this risk by instructing the whole team to the FlowLight, and by allowing the participants to get used to the FlowLight for a whole work week after the installation before we started to evaluate its effects.

Construct Validity. One of the major threats to construct validity of our research results on sensing interruptibility is the use of the experience sampling technique applied to collect self-reports of interruptibility. To collect the ground truth data for participants' interruptibility, we interrupted them and asked them to rate it. Potentially, this interruption and thereby induced context switch could have led to wrong answers of the prompts. To mitigate this risk, we developed the pop-up to collect the self-reports to be as nonintrusive as possible, e.g. by ensuring that only one click was needed to answer a prompt. Additionally, the participants had the possibility to postpone or skip the prompts if they were not able to answer them. Further, prior work has shown that short interruptions are less disruptive than longer ones and additionally our participants stated that they did not feel particularly disrupted by the prompts.

Another, related potential threat to the construct validity is that participants might not have answered the self-reports consistently throughout the study period, e.g. using different scales for different days, in particular in the two-week study (*study 3*), or using different scales per participant, which might apply to all three studies (*study 1*, *study 2*, and *study 3*). We tried to mitigate this risk by explaining the prompts for the self-reports thoroughly and in person

prior to starting the study, and by applying normalization techniques of the interruptibility ratings across participants.

A threat to the construct validity in the evaluation of the FlowLight might lie in the self-reporting of interruptions. Participants might not have reported all interruptions during the two weeks of interruption logging, and the work patterns before and after the installation of the FlowLight might have been significantly different, which makes it more difficult to compare the effect of the FlowLight. We tried to mitigate these risks by only including the logs of participants who logged interruptions for more than three days before and three days after the installation, and by regularly reminding them about the logging.

1.5 Challenges

In the following, we point out the major challenges we faced throughout our research, including the trade-off between accuracy and privacy, a feeling of being observed induced by disclosing a person's interruptibility to co-workers, the usage of biometric sensors in the field, and the recruiting of study participants.

Accuracy vs. Privacy. One of the major challenges for our research is the sensitivity of the collected data and the associated privacy concerns. While more and fine-grained data, such as knowing each web site visited or the sleep duration, provides valuable information to sense interruptibility more accurately, it also reveals much about a person's daily life. In our research, we faced the trade-off between obtaining a high accuracy of the interruptibility measurement and limiting the richness and quantity of the collected data to reduce privacy concerns. While we treated the collected data confidentially and anonymized all identifying data, we also examined which and how much data is necessary to still achieve a high accuracy. For example, for the FlowLight, we were able to build an interruptibility metric solely based on keystroke and mouse input. The developed metric was accurate enough to reduce interruption cost successfully, and at the same time required only a small set of data types. To make it more accurate, later versions of the tool offered the possibility to provide more data

voluntarily by adding certain applications to a blacklist. Applications on the blacklist were not taken into account to calculate the FlowLight's interruptibility status, which means that it would not switch to *busy* or *do not disturb* during elevated activity in these applications. In particular, many participants did not want the status to be changed to *do not disturb* during the use of an instant messaging client. To further preserve privacy of the leveraged data, all input data of the interruptibility measurement was stored locally on the participants' computers, leaving the user in full control of his or her data.

Feeling Observed. A further challenge was that some of the FlowLight's users sometimes felt observed, as they were indicating their interruptibility to their colleagues. Participants often felt that having a red or even pulsating red light is linked to focus or productivity, and vice versa some participants felt that a green light might be understood as being unproductive or distracted. We addressed this fear of being observed and perceived as less productive than others by limiting the duration a participant's light would be red or pulsating red to 9% of the participant's day. While there might be slight variance in the actual duration per participant across days, since the threshold was based on the participant's interaction during previous days, overall it is pretty stable across days and participants. We determined the threshold based on early user feedback in pilot studies. We further instructed participants that a green light would not at all indicate that someone was not working productively, but just that at these times, the cost of interruptions are lower compared to the times of a red / pulsating red light. We further stressed the importance of offering enough times where one is available for questions or discussions to colleagues to increase the team's productivity.

Using Biometric Sensors in the Field. Biometric sensors come in different form factors (e.g chest straps or wrist bands) with different levels of comfort, ease of use, reliability, data richness and accuracy, and access to the data [Hänsel et al., 2018]. While we were interested in evaluating a wide range of biometric measurements in the field, we needed to consider certain criteria to select sensors

which were feasible to use in our studies. Some sensors provide rich data and are applicable in controlled lab experiments, but they might not be feasible to use on a daily basis in a field study of several weeks due to lack of comfort or limited battery life. As an example, EEG sensors provide insightful data which have been linked to cognitive load and attention, and are therefore most likely also linked to interruptibility. However, many EEG sensors are rather invasive as they come in the form of many electrodes to be placed on the head. In our research, we were able to use a little-invasive EEG device, the Neurosky Mindband in the form of a headband with only three electrodes. However, the usage of this sensor was only possible in the short lab and field study (*study 1* and *study 2*), since the lack of comfort and battery life limited its usage to a few hours. For the two-week field study *(study 3)* we required sensors with minimal maintenance effort to be used on a daily basis, while still providing valuable and accurate data. These considerations influenced our decision to use the Fitbit Charge 2 and the Polar H7. We believe as sensing technologies improve in the future, it will be possible to evaluate and use additional kinds of sensors in a field context.

Finding Study Participants. Finding study participants who were willing to wear sensors for our study for up to two weeks was one of the challenging and time consuming tasks of the work presented in this thesis. The challenge was that on one hand, we had certain technical constraints, in particular that our monitoring tool only runs on the Windows operating system; and on the other hand, participants had to be willing to wear several biometric sensors and share their data with us. In addition, due to the sensitivity of some of the collected data, it had to be acceptable for the company that employees would share their data with us. We therefore spent a substantial amount of time to recruit participants, e.g. by preparing invitations to our studies and cultivating connections with companies.

We also observed that providing insights about personal habits based on the collected biometric data and computer interaction data, was a strong motivator for many to participate in the study. Our research group therefore put effort

into providing visualizations of the data as well presenting the aggregated results to the participants after we analyzed it.

1.6 Opportunities and Future Work

We want to point out multiple opportunities for future research, including applications of the interruptibility model, increasing its accuracy and applicability, and investigating new ways to reduce interruption cost for both the interruptee and interrupter.

Interruptibility Model in Practice. Based on the findings from *RQ 1* and *RQ 2*, an obvious next step is to use the determined interruptibility model from *RQ 1* for the FlowLight (*RQ 2*). While the simple algorithm of the FlowLight was already accurate enough to reduce interruption cost, the model developed in *RQ 1b* is based on more detailed data from computer interactions and also biometric measurements and thus can be more accurate for a broader range of activities. These activities include situations where a knowledge worker is highly focused but not interacting actively with the computer, such as times spent understanding source code or reading a document, or working with pen and paper. The general interruptibility model trained across multiple participants can be used to achieve a high accuracy without the need of an initial training phase which solves the cold-start-problem. In addition, we could incorporate a feature to collect interruptibility ratings for a few days to train it to the individual and further increase accuracy. Given the variety of sensors used in our study from more to less physically and privacy invasive, we can further take into account the users' preferences on which sensors and features to use. For instance, users might prefer using a biometric tracker over a computer interaction tracker or vice versa.

While the major focus of the FlowLight is to reduce in-person interruptions, the interruptibility model could also be used to support the handling of computer-based interruptions, such as emails or instant message notifications. Most existing work has focused on finding naturally occurring task boundaries to display

interruptions (e.g. [Iqbal and Bailey, 2008]), while our model would provide a more continuous interruptibility measurement and only display interruptions when the interruptibility level is above a certain threshold. Such a continuous measurement would allow deferring computer-based interruptions not only to task boundaries but to other times when interruption cost are particularly low, and addressing computer-based and in-person interruptions at the same time.

External interruptions, either coming from in-person interruptions or from notifications at the computer, only make up half of the interruptions that a knowledge worker experiences in a day. The other half stems from self-interruptions, such as switching to check a news website [Czerwinski et al., 2004]. These self-interruptions can also have a big impact on the developer's focus and performance. Since our interruptibility measurement is presumably also related to focus, we might be able to use it to reduce the cost of self-interruptions and better support developers in their work. For instance, by automatically detecting when a knowledge worker's focus is decreasing, we might be able to intervene, e.g. by reducing distracting content on the screen that might cause self-interruptions or by suggesting to take a break. Furthermore, by knowing when a knowledge worker is more or less focused during the day, we might be able to optimize the work day by scheduling highly demanding tasks during times of high focus.

Increasing the Applicability and Accuracy. In our work, we focused on software developers as one coherent group of knowledge workers. We see great potential in targeting the sensors and features towards the specific kind of work of the users' job roles, possibly increasing the accuracy of the model. As an example, in our work, we added the time spent in activities related to software development as a feature for our interruptibility model. Future research could investigate meaningful features for other job areas, e.g. designers or other engineers.

Further, we focused on sensing interruptibility while a knowledge worker is working with the computer or close by. Yet, to cover a broader range of activities, such as having discussions with colleagues or while interacting with other devices, integrating additional data sources can potentially increase our

model's accuracy. Data sources such as interactions with mobile devices, audio or location logs can be valuable and have already been explored in sensing interruptibility (see [Turner et al., 2015] for a review). Obviously, further data sources also add privacy concerns and future research could focus on evaluating costs, benefits and predictive power of each data source and integrate valuable sources into one holistic model.

Our research serves as a good starting point evaluating a variety of biometric sensors in office workplaces. As new and improved biometric sensing technologies are emerging frequently given the fast advances in their development, it might be possible to collect and use additional biometric data in the field and over a prolonged time period. As an example, using attentive states detected with EEG data can be valuable additional features in the prediction of interruptibility, increasing its accuracy even more. Additionally, the capabilities of regular devices to sense biometric features are growing, e.g. web cams have been used to predict emotions based on facial expressions [Bahreini et al., 2016] and cognitive load based on pupil dilation [Samara et al., 2017].

Further Ways to Reduce Interruption Cost. Our approach helps to protect the interruptee from interruptions at inopportune moments. However, after the interruption, it might still take some time to refocus and remember the context of the suspended task. A potential research direction could therefore be to leverage computer interaction data such as recently used programs, files, or websites to summarize or highlight relevant contextual information of the suspended task when the user returns from an interruption. Such an aid can potentially reduce the resumption lag—the time to resume the primary task—and thus decrease the cost of interruptions even more [Iqbal and Horvitz, 2007, Rule et al., 2015].

Further, researchers have shown that additional to the interruptee's interruptibility state, other characteristics such as the urgency, importance and context of both the primary and interrupting tasks are important factors to find optimal moments for interruptions [Arroyo and Selker, 2011, Grandhi and Jones, 2010]. In our approach, the interruptee and interrupter are aware of the interruptee's interruptibility state, but still need to assess these additional factors by themselves

to decide when to address the interruption. Future research might integrate our interruptibility model with these factors to support the decision making process, e.g. by mediating interruptions directly [Kobayashi et al., 2015, Iqbal and Bailey, 2008], or by displaying interruptions at opportune moments along with additional information on these factors [Grandhi and Jones, 2015].

1.7 Background and Related Work

The main objective of this work is to support knowledge workers and reduce their interruption cost. To achieve this objective, we aim to measure interruptibility at any given point in time and to support better handling of in-person interruptions by providing awareness on a person's interruptibility to co-workers. In this section, we will first provide an overview of related work on interruptions in the workplace, before we discuss studies on the sensing of interruptibility and finally approaches to better handle interruptions.

1.7.1 Interruptions in the Workplace

Several observational studies found that a typical work day of knowledge workers is highly fragmented [Czerwinski et al., 2004, González and Mark, 2004]. They get interrupted 25 times a day on average, half of these interruptions being self-initiated and the other half being caused from external persons or systems, e.g. personal visits, email notifications or phone calls [González and Mark, 2004]. For these interruptions, a knowledge worker spends about 15-20 minutes per interruption and overall a total amount of 15-20% of the workday on handling them [van Solingen et al., 1998]. Among these interruptions, the ones that take the longest are personal visits from colleagues (ranging from 24 minutes up to 4 hours) [Sykes, 2011].

While interruptions are necessary in a collaborative work environment and can increase productivity [van Solingen et al., 1998], they can also have a variety of negative effects as multiple studies have shown. These negative effects range from increased time needed to complete a task to a higher number of errors and

increased annoyance. The interruption cost is particularly high if interruptions happen at inopportune moments, e.g. when a person is highly engaged in a task [Bailey and Konstan, 2006]. Often, knowledge workers do not even go back to their suspended task directly after an interruption occurred. On average, they engage in two other tasks before resuming the suspended task, and 27% of task suspensions result in more than two hours before resumption, which increases the overall time needed to make progress [Mark et al., 2005, Iqbal and Horvitz, 2007]. Another study found no increase in the time needed to complete an interrupted task, but more stress and frustration since the participants wanted to compensate for the interruptions by working faster [Mark et al., 2008]. It can be speculated, that this stems from the Zeigarnik effect: a strong motivation to work with heightened efficiency after being interrupted [Zeigarnik, 1927, Brehmer et al., 2012].

Not all interruptions are equally disruptive. Studies investigating the disruptiveness of different kinds of interruptions found the interruption duration, the difficulty of the interrupting task, the relevance of the interrupting task to the current task, the interruption moment and the interruption frequency to be important factors [Bailey and Iqbal, 2008, Czerwinski et al., 2000, Monk et al., 2008]. Further, interruptions are less disruptive if the interrupted person has the possibility to choose a suitable moment for an interruption as opposed to having to respond immediately [McFarlane, 2002]. Times with low perceived cognitive load have been shown to be suitable moments for interruptions, as several researchers showed (e.g., [Chen et al., 2007, Bailey and Iqbal, 2008]). The experienced amount of cognitive load can vary widely, even between individuals working on the same task, since a person's age, personality traits, or prior knowledge can also influence cognitive load. Cognitive load generally refers to the total amount of required mental effort to perform a task and is composed of intrinsic, extrinsic and germane load [Sweller, 2011]. The intrinsic cognitive load is posed by the intrinsic characteristics of the current task, the extraneous cognitive load by the form in which the task is presented and the germane load depends on the effort that is needed to process the information at hand [Sweller, 1994].

In our work, we want to reduce the overall cost of interruptions for knowledge workers. In particular, we focus on in-person interruptions, since these belong to the most time consuming and disruptive kind of interruptions due to their immediate nature [Sykes, 2011, McFarlane, 2002].

1.7.2 Sensing Interruptibility

When a co-worker is asked to assess a colleague's interruptibility into five states—from highly interruptible to highly non-interruptible—the assessment is difficult and only slightly better than chance as Fogarty et al. found in their study [Fogarty et al., 2005a]. These assessments are generally based on cues of the colleague's social and task engagement, such as an open door, the colleague talking to someone or the use of the computer keyboard. However, especially in today's globally distributed work environments, this context information is often not easily accessible to remote colleagues.

Researchers have examined various data sources to determine a person's interruptibility automatically and with high accuracy. Most prior work thereby focused on contextual sensors that monitor interactions of a person with its environment, e.g. by recording audio, video, interactions with mouse, keyboard, or applications (e.g. [Horvitz et al., 2002, Begole et al., 2004]. Few researchers further examined the use of biometric sensors, which measure reactions of the body to external stimuli, which are potentially related to interruptibility [Mathan et al., 2007,Chen et al., 2007]. Biometric sensors have been linked to mental states such as cognitive load, stress, or emotions in various psychological and psychophysiological studies, theories and concepts (e.g. [Grimes et al., 2008, Boucsein, 2012]).

In the following, we will discuss various studies and approaches that have used contextual and biometric sensors to measure cognitive and emotional states, and also interruptibility. In our study, we build upon this research and extend it by combining a variety of different sensors and studying them in the field to measure interruptibility. To the best of our knowledge, our approach is the first to conduct studies in knowledge workers' workplaces and combine and compare computer

interaction sensors, which is one type of contextual sensors, and biometric sensors for this purpose.

Contextual Sensors

To measure a knowledge worker's interruptibility, several researchers have leveraged information on a person's interaction with the environment and devices. Prior work studied a wide variety of data sources to capture context, ranging from audio and video recordings, over calendar or network connection data to computer interaction data, which mainly consists of mouse, keyboard, and application usage data.

Some researchers followed a wizard-of-oz approach to determine interruptibility based on contextual information. For instance, Hudson et al. classified interruptibility by simulating sensors and manually coding features from audio and video recordings, such as the number of people present, who was speaking, or whether the phone was on the hook [Hudson et al., 2003]. Iqbal et al. used task characteristics, such as the next subtask's difficulty, carry over of data across boundaries, and the percentage of parent task completion, to predict the cost of interruptions in terms of the resumption lag—the time needed to resume the primary task [Iqbal and Bailey, 2006].

In contrast to these sensing approaches which require manual coding, various researchers investigated how contextual data can be leveraged to automatically measure interruptibility. As an example, Fogarty et al. focused on a specific type of computer interaction data, namely IDE interaction data, to measure interruptibility during software development tasks [Fogarty et al., 2005b]. Another body of research measured interruptibility by combining more general computer interaction data such as keyboard and mouse interaction with further contextual data such as location, speech, calendar, time, presence or network data [Begole et al., 2004, Lai et al., 2003, Fogarty et al., 2004, Horvitz et al., 2004, Kapoor and Horvitz, 2007, Kapoor and Horvitz, 2008, Horvitz et al., 2002]. Several researchers have focused on the use of contextual data to automatically identify naturally occurring breakpoints during task execution while working on the computer, which are opportune moments for interruptions since cognitive load

drops at these moments [Borst et al., 2015, Bailey and Iqbal, 2008]. The studies on sensing breakpoints mainly investigated computer interaction features such as the frequency of window switches and ranged from a few hours [Tanaka and Fujita, 2011, Iqbal and Bailey, 2008] to 2 weeks [Nair et al., 2005].

To summarize, these approaches demonstrate that different combinations of contextual sensors can be used to measure interruptibility in specific contexts and tasks. It is not yet known which of these sensors are most accurate in sensing interruptibility continuously in the field and how they compare to biometric sensors. In our work, we want to evaluate contextual sensors, in particular a wide variety of computer interaction sensors, in their ability to continuously and accurately measure knowledge workers' interruptibility, and compare and combine them to biometric sensors.

Biometric Sensors

Background on Biometric Sensing. Biometric sensors have been used in a variety of studies to detect aspects related to cognitive load, stress, emotions, or health. Since perceived cognitive load and emotions are highly individual, biometric sensors bear great potential to capture these individual differences. The existing studies can be differentiated by the type of biometric sensor and data being used, in particular sensors measuring the activity of the brain, eye, heart, skin, and the body's movement. The most studied sensors are EEG, eye tracking systems, electrocardiographs (ECG) or blood volume pulse (BVP) sensors, EDA and body temperature sensors.

Brain. EEG measures the aggregated electrical activity of the brain, which is caused when neurons fire. Different studies showed that certain frequency bands in the EEG data, called Alpha, Beta, Gamma, Delta and Theta can be linked to cognitive states, such as being focused, relaxed, or dreaming [Berger, 1929]. For instance, Gevins et al. found that an increase in theta activity and a decrease in alpha activity can be linked to an increase in memory load [Gevins et al., 1998]. Several studies also used EEG devices to classify mental tasks or states of cognitive load [Grimes et al., 2008, Lemaire, 1996].

Eye. Eye trackers use the reflection of infrared light from the eyes to calculate the position of the visual focus and the pupil size. Interesting features are the pupil size, fixation duration or number of saccades. Particularly the pupil size (e.g. the peak amplitude of the pupil diameter) is an indicator for memory load or processing load, and varies with task difficulty [Beatty, 1982]. Researchers showed that more difficult tasks demand longer processing time, induce higher subjective ratings of cognitive load and evoke greater pupillary response at salient subtasks [Iqbal et al., 2004b].

Heart. For measuring the activity of the heart, such as the heart rate (HR), either an ECG or BVP sensor can be used. ECG sensors measure the electrical activity of the heart using electrodes which are placed on the chest. BVP sensors emit light which is absorbed by the oxyhemoglobin in the blood. The part of the light which is scattered back can be detected with a photodiode. Both ECG and BVP have been used to measure the cognitive load [Haapalainen et al., 2010, Peper et al., 2007]. As a further important feature related to the heart's activity is the heart rate variability (HRV), which can be derived from the ECG signal and was found to be related to stress [Hjortskov et al., 2004].

Skin. Electrodermal activity (EDA) refers to the skin conductivity that varies with sweating activity and can be measured by applying a small current with two electrodes. EDA has been linked to arousal, attention, emotional states, stress and anxiety [Boucsein, 2012]. In a study on text reading and arithmetic tasks imposing multiple cognitive load levels, a strong link between cognitive load and EDA was found [Nourbakhsh et al., 2012]. In addition to EDA, skin and body temperature have also been linked to cognitive load, emotions, as well as stress. For example, skin temperature was demonstrated to be different in response to anger and fear [Collet et al., 1997] and the heat flux has been shown to be able to predict cognitive load [Haapalainen et al., 2010]. Vinkers et al. recently confirmed indications that stress influences body temperature in humans and found that body temperature rises with increasing stress [Vinkers et al., 2013].

Physical activity. Physical activity is defined as "all bodily actions produced by the contraction of skeletal muscle that increase energy expenditure above basal level" and is mostly measured using accelerometer data, sometimes accompanied

by HR data [Butte et al., 2012]. Sensors to measure physical activity mostly come in the form of wrist-bands or small devices attached to the hip or a leg. Recent advances of such sensors allow to measure and research physical activity during the whole 24-hour cycle, spanning over bodily activities at work, during leisure time and also sleep duration and quality at night [Rosenberger et al., 2016]. As an example, sleep has been shown to have a big impact on productivity and mood [Rosekind et al., 2010, Vidaček et al., 1986, Mark et al., 2016a].

Sensing Interruptibility with Biometric Sensors. While biometric sensors have been used in various studies on measuring cognitive load or emotions, very few studies examined the use of this type of sensors to measure interruptibility. Most of these studies focused on one specific sensor type. For example, Kramer gathered EEG data during one hour of US military training and succeeded in classifying interruptibility based on labels gathered retrospectively [Mathan et al., 2007]. Bailey and Iqbal focused on eye tracking to measure interruptibility [Bailey and Iqbal, 2008]. In particular, they compared mental workload during different hierarchic levels of task boundaries and were able to show that mental workload dropped at high-level task boundaries. Chen et al. used HRV and electromyography (EMG) to measure interruptibility. They conducted an experience sampling study in which participants solved short tasks with varying difficulty. They used the HRV measurement as an indicator for cognitive load and the EMG to detect muscle activity and calculated interruptibility using linear regression [Chen et al., 2007]. Furthermore, accelerometer data has been used in several studies to detect physical activity and to show that interruptions are better delivered during moments recognized as activity transitions, e.g. when walking to another location [Ho and Intille, 2005, Fisher and Simmons, 2011, Komuro et al., 2017].

The existing studies provide initial evidence of the potential of biometric sensors to measure interruptibility for selected tasks. In our research, we extend this research and investigate whether we can use such sensors to measure interruptibility not just in the lab and short tasks, but in the field while knowledge workers perform their work as usual.

1.7.3 Supporting Interruption Handling

Due to the high cost of certain interruptions, several approaches and methods have been proposed to improve the handling of interruptions and reduce their cost. These approaches range from simple manual strategies that knowledge workers use in their every-day life to more advanced and automatic systems developed by various researchers.

Several researchers studied how knowledge workers cope with interruptions. For instance, González and Mark observed knowledge workers and investigated strategies they apply to manage different activities in order to remind themselves of relevant information and goals, as they usually experience a high level of discontinuity in the execution of their activities [González and Mark, 2004]. They found that knowledge workers use post-it notes, print-outs of email messages and planners to manage their variety of tasks and cope with the fragmented nature of their work. Another common strategy to deal with unwanted interruptions is the usage of earphones or ear buds, to either signal that one does not want to be disturbed or to tune out distractions [Sykes, 2011]. Rather than looking at the prevention of interruptions, Parnin and Rugaber investigated strategies to better deal with interruptions, in particular, resumption after an interruption during a programming task [Parnin and Rugaber, 2011]. They found that programmers went back to the last edit, navigated through code or looked at other task specific information in the bug tracking tool or the revision history to rebuild their context and resume the current task.

In addition to the strategies that knowledge workers came up with, researchers have also developed approaches to support interruption handling. The most prominent techniques in this domain are strategies to defer interruptions to breakpoints and to provide awareness on the interruptibility of a knowledge worker automatically and continuously to co-workers.

Postponing Interruptions to Breakpoints. Several researchers developed tools to mediate interruptions by postponing them to more opportune moments. Most of these approaches implement the defer-to-boundary policy that aims at finding natural breakpoints during work and delaying interruptions to these breakpoints

instead of displaying them immediately [Iqbal et al., 2004a]. This idea is based on studies that found that the cognitive load drops at task boundaries, and that interruptions at lower cognitive load are less harmful [Bailey and Iqbal, 2008,Borst et al., 2015]. Researchers predominantly applied the defer-to-boundary strategy to computer-based interruptions, such as notifications from incoming emails or instant messages. For instance, a decision-rule-based software developed by Arroyo and Selker delivers unrelated instant messages only at times of context switches that were determined based on mouse movement and window switching. In their study they achieved a five times higher answer rate to messages compared to non-mediated message delivery [Arroyo and Selker, 2011]. Iqbal and Bailey developed a system that implements the defer-to-breakpoint policy to reschedule notifications to more opportune moments [Iqbal and Bailey, 2008]. In a study, they found that notifications delivered at breakpoints caused less frustration and a shorter reaction time compared to notifications which were delivered immediately.

Indicating Interruptibility. Another strategy to better handle interruptions is to indicate a knowledge worker's interruptibility continuously to potential interrupters. While knowledge workers already use simple and manual indicators such as headphones [Sykes, 2011] or manual busy lights [Embrava, 2016], researchers have also developed applications to indicate interruptibility based on automatic interruptibility measurements.

Most prior work that designed availability indicators, created contact-list style tools installed on the computer along with information on a person's interruptibility. Examples are Connexus, Lilsys, and MyVine. Connexus integrates awareness information, instant messaging and other communication channels and indicates availability to potential interrupters [Tang et al., 2001]. Lilsys and MyVine extend this approach mainly by adding further data sources to improve the accuracy of the interruptibility measurement [Begole et al., 2004, Fogarty et al., 2004]. Evaluations of these tools indicate that these contact-list style tools did not reduce in-person interruptions, but could show a qualitative improvement in interruption awareness.

As in-person interruptions are the most common and expensive interruption in the workplace, in particular due to their immediate nature and disruptiveness [Sykes, 2011], researchers have also tried to address in-person interruptions by indicating interruptibility to co-workers, but more research is needed to assess the effects of such approaches.

The research most similar to the FlowLight presented in this thesis is by Milan et al. In their work, they developed an automatic interruptibility indicator based on the cost of interruptions during certain automatically detected activities (e.g. email or being in a meeting) using audio, video and computer interaction data [Bjelica et al., 2011]. To evaluate their approach, they conducted a small user study with one participant in an office environment and over the course of a single workday per indication modality. In their study, they investigated the effects of two modalities of the interruptibility indicator: a busy flag (small light placed on desk), and ambient lighting effects. Both modalities for indicating the status decreased the number of interruptions, yet the ambient lighting effect had a bigger effect.

In our work, we combine an automatic interruptibility measure based on computer interaction with a physical indicator in the form of a traffic-light like LED placed on a knowledge worker's desk. We conducted a large-scale and long-term user study to investigate the effects of such an automatic indicator in office workplaces and the interaction behavior of knowledge workers.

1.8 Summary and Contribution

Our findings from *study 1*, *study 2*, and *study 3* support our ***Hypothesis H1*** by providing strong evidence that a knowledge worker's interruptibility can be measured automatically with high accuracy based on computer interaction and biometric data. The model based solely on computer interaction data achieved the highest accuracy for times the knowledge worker spent at the computer. Biometric data can improve the accuracy and is especially useful at times spent away from the computer. Further, our interruptibility model is able to achieve

a high accuracy both for individual but also general models, which solves the cold-start problem.

With the FlowLight, we built and evaluated one approach to successfully reduce interruption cost in the office. This supports ***Hypothesis H2***, which states that an automatic interruptibility indicator in the form of a physical light can reduce the cost of in-person interruptions in an office work environment. The physical traffic-light like LED lamp targets in-person interruptions, reducing them by 46%, increasing the awareness on the potential harm of interruptions, and improving the perceived productivity.

Our work makes the following contributions:

- we present findings from a one-hour lab and two-hour field study on the use of biometric sensors that demonstrate their feasibility to predict interruptibility;
- we present findings from a two-week field study on the use of computer interaction and biometric sensors showing that it is possible to sense interruptibility in office workplaces continuously with high accuracy;
- we present and discuss the development and evaluation of a model to sense interruptibility, also assessing the best combination of features and optimal time window selection;
- we present the FlowLight, an approach to automatically sense and indicate a knowledge worker's interruptibility in the form of a physical traffic-light like LED lamp;
- we present results from a large-scale field study showing that the FlowLight successfully reduces the cost of in-person interruptions in office workplaces.

1.9 Thesis Roadmap

The remainder of this thesis consists of four chapters, each published at an internationally renowned, peer-reviewed conference. An overview of all publications

is given in Figure 1.2.

Chapter 2 addresses *RQ 1a*. Using data collected in a one-hour lab and a two-hour field study, we investigated the feasibility of interruptibility prediction with a variety of biometric sensors. These studies and their findings provide initial evidence that it is possible to sense interruptibility in office workplaces continuously and with high accuracy.

Chapter 3 addresses *RQ 1b*. With a two-week field study on interruptibility sensing, we extended our research described in Chapter 2. The study investigates a broader range of sensors in the field, including both biometric and computer interaction sensors, and provides a comparison of the accuracy of various sensors and features to sense interruptibility in office workplaces.

Chapter 4 addresses *RQ 2*. Using an automatic measure of interruptibility based on mouse and keystroke data, we developed the FlowLight, an automatic physical interruptibility indicator light which reduces in-person interruption cost. Chapter 4 discusses the approach and presents the results of a large-scale field study, showing the positive effect of the FlowLight on interruption cost.

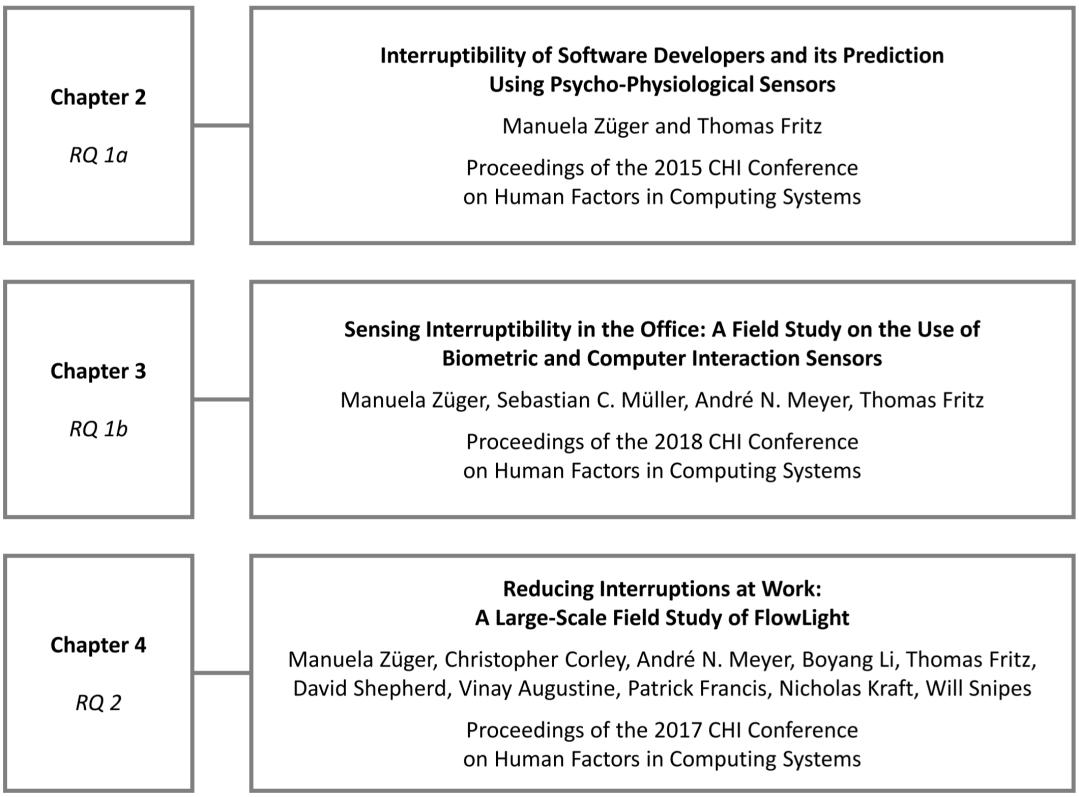


Figure 1.2: Thesis Roadmap.

2

Interruptibility of Software Developers and its Prediction Using Psycho-Physiological Sensors

Manuela Züger and Thomas Fritz
Published at the 2015 CHI Conference on Human Factors in Computing Systems
Contribution: Study design, data collection, part of data analysis, and paper writing

Abstract

Interruptions of knowledge workers are common and can cause a high cost if they happen at inopportune moments. With recent advances in psycho-physiological sensors and their link to cognitive and emotional states, we are interested whether

such sensors might be used to measure interruptibility of a knowledge worker. In a lab and a field study with a total of twenty software developers, we examined the use of psycho-physiological sensors in a real-world context. The results show that a Naïve Bayes classifier based on psycho-physiological features can be used to automatically assess states of a knowledge worker's interruptibility with high accuracy in the lab as well as in the field. Our results demonstrate the potential of these sensors to avoid expensive interruptions in a real-world context. Based on brief interviews, we further discuss the usage of such an interruptibility measure and interruption support for software developers.

2.1 Introduction

In office work environments, interruptions by co-workers, emails or instant messages are common. While some of these interruptions are desired, others might incur a high cost, including long resumption lags, slower performance, negative emotions and an increase in errors due to the interruption happening at inopportune moments [Czerwinski et al., 2004, Bailey and Konstan, 2006]. When a co-worker is asked to assess a colleague's interruptibility into five states—from highly interruptible to highly non-interruptible—the assessment is difficult and only slightly better than chance as Fogarty et al. found in their study [Fogarty et al., 2005a]. In addition, these assessments are generally based on cues of the colleague's social and task engagement, such as an open door, the colleague talking to someone or the use of the computer keyboard. However, especially in today's globally distributed work environments, this context information is often not available.

To avoid the high costs that interruptions can cause on knowledge workers, researchers have looked at automatically identifying good and bad moments for interruptions and computing a measure for a worker's interruptibility. Such an automatic interruptibility measure can then be used to better coordinate interruptions by, for instance, providing visual cues or postponing them to a more suitable moment [McFarlane, 2002]. Prior work examined the use of context-aware sensors to gather information, such as audio and video streams, keyboard

and mouse interaction, or task characteristics (e.g., [Fogarty et al., 2005a,Fogarty et al., 2005b,Iqbal and Bailey, 2006]).

With recent advances in psycho-physiological (aka. biometric) sensor technology, researchers have also started to investigate their use to assess a person's interruptibility in controlled environments and on small tasks. Psycho-physiological sensors have the advantage of providing more flexibility without being bound to a specific task, computer or location and are increasingly less invasive. Previous studies have shown that psycho-physiological features, such as electrodermal activity (EDA), heart rate or electroencephalographs (EEG), can be used to measure cognitive and emotional states (e.g., [Berger, 1929,Haapalainen et al., 2010]). Under the assumption that moments of high cognitive load or stress correlate with low interruptibility, studies have examined, for instance, the use of psycho-physiological sensors to calculate interruptibility from EEG data during a military exercise [Mathan et al., 2007] or from a combination of heart rate variability and muscle activity on varying small tasks, such as word puzzles, mental arithmetic or racing games [Chen et al., 2007].

In our work, we aim to investigate the use of a combination of psycho-physiological sensors to automatically identify the interruptibility of a knowledge worker in a real-world working context. We build upon and extend the findings of previous research in the area by contributing two studies with software developers wearing psycho-physiological sensors: a lab study with participants working on real-world development tasks in a controlled environment, and a field study with participants working on their own tasks in their real-world office environments. In our analysis we focus on the use of psycho-physiological sensors to automatically infer the interruptibility of knowledge workers. In addition, we investigate the correlation between interruptibility, mental load and interruption lag and look at which tool support for interruptions developers desire.

The results of our studies provide evidence that psycho-physiological sensors can be used to classify the interruptibility of a software developer with high accuracy (91.5% for the lab and 78.6% for the field study). Our results also show that we can build classifiers with high accuracy for a more fine-grained set of five states of interruptibility and that psycho-physiological data is very sensitive to

individuals. In addition to these results, we provide evidence for the correlations between interruptibility, mental load and interruption lag and discuss potential tool support for software developers.

2.2 Related Work

2.2.1 Interruptibility with Context-Aware Sensors

The greater part of related work that developed an interruptibility measure used context-aware sensors, such as audio and video streams, keyboard and mouse actions, active window information, table pressure, or task characteristics. For instance, Hudson et al. were able to classify interruptibility into two states (least interruptible vs. all other states) with 78% accuracy by simulating sensors and manually coding features from audio and video recordings, such as the number of people present, who was speaking, or whether the phone was on the hook [Hudson et al., 2003]. Fogarty et al. also simulated sensors by manually encoding mouse and keyboard actions, such as highlighting a line or editing code. They measured interruptibility in terms of the interruption lag—the time between the notification and the start of an interruption—during coding tasks and achieved 72% accuracy for two state interruptibility classification (interruptible and engaged) [Fogarty et al., 2005b].

Using a pressure sensor sheet on the desk, Tani et al. were able to achieve a similar two state interruptibility classification accuracy on typing and mouse operation with easy and hard phases [Tani and Yamada, 2013].

Different to these, Iqpbal et al. used task characteristics, such as the next subtask's difficulty, carry over of data across boundaries, and the percentage of parent task completion, to predict the cost of interruptions. They measured the cost based on resumption lag—the time needed to resume the primary task [Iqbal and Bailey, 2006].

Ho et al. developed a context-aware mobile computing device to automatically detect activity transitions using accelerometers for measuring interruptibility.

Their results show that messages delivered at activity transitions are better received than those delivered at random times [Ho and Intille, 2005].

Instead of context-aware sensors, our work uses biometric sensors to assess interruptibility. Especially with the recent advances and the development of low-cost biometric sensors with low-invasive form factors, a possible advantage is that these sensors are body-worn and not limited to laboratories, certain working environments or a specific software platform.

2.2.2 Biometric Sensors

An extensive amount of research investigates the link between psycho-physiological sensors and different cognitive and emotional states, such as high cognitive load or stress.

The most studied sensors are electroencephalographs (EEG), eye tracking systems, sensors for electrocardiogram (ECG) or blood volume pulse (BVP), sensors measuring the electrodermal activity (EDA) and body temperature sensors.

EEG. EEG measures the aggregated electrical activity of the brain, which is caused when neurons fire. Different studies showed that certain frequency bands in the EEG data, called Alpha, Beta, Gamma, Delta and Theta can be linked to cognitive states, such as being focused, relaxed, or dreaming [Berger, 1929]. For instance, Gevins et al. found that an increase in theta activity and a decrease in alpha activity can be linked to an increase in memory load [Gevins et al., 1998]. Kramer linked an increase in beta and decreases in alpha and theta to an increase in task engagement [Kramer, 1991]. Several studies also used EEG devices to classify mental tasks or states of cognitive load [Lee and Tan, 2006, Grimes et al., 2008].

Eye Tracking. Eye trackers use the reflection of infrared light from the eyes to calculate the position of the visual focus and the pupil size. Interesting features are the pupil size, fixation duration or number of saccades. Particularly the pupil size (e.g. the peak amplitude of the pupil diameter) is an indicator for

memory load or processing load, and varies with task difficulty [Beatty, 1982]. Researchers showed that more difficult tasks demand longer processing time, induce higher subjective ratings of cognitive load and evoke greater pupillary response at salient subtasks [Iqbal et al., 2004b]. Fixation durations and number of saccades are suitable to assess the designs of user interfaces [Jacob and Karn, 2003].

ECG and BVP. For measuring the activity of the heart, either an ECG or BVP sensor can be used. ECG sensors measure the electrical activity of the heart using electrodes which are placed on the chest. BVP sensors, or photoplethysmographs, emit light which is absorbed by the oxyhemoglobin in the blood. The part of the light which is scattered back can be detected with a photodiode. ECG devices are more exact, but also more cumbersome to wear, therefore we chose a BVP sensor for our study. BVP can serve as a direct indicator for cognitive load [Peper et al., 2007], and can additionally be used to compute interbeat interval (IBI) and heart rate (HR).

EDA. EDA represents the skin conductivity that varies with sweating activity and can be measured by applying a small current with two electrodes. EDA has been linked to arousal, attention, emotional states, stress and anxiety [Boucsein, 2012]. In a study on text reading and arithmetic tasks imposing multiple cognitive load levels, a strong link between cognitive load and EDA was found [Nourbakhsh et al., 2012].

Body temperature. Body temperature is influenced by emotions as well as stress. Vinkers et al. recently confirmed indications that stress influences body temperature in humans and found that body temperature rises with increasing stress [Vinkers et al., 2013]. In a study about autonomic nervous system response patterns evoked by emotions, skin temperature was demonstrated to be different in response to anger and fear [Collet et al., 1997].

Sensor combinations. Studies also applied multiple biometric sensors to measure cognitive load, task difficulty, task engagement and other cognitive states.

For instance, Wilson analyzed mental workload in pilots during flight with multiple measures [Wilson, 2002]. In prior work, we combined EDA, EEG and eye tracking to assess task difficulty in simple code comprehension tasks and found that a combination of all sensors to classify a new task achieved the highest accuracy of 84%. Haapalainen et al. combined an eye tracker, a heart rate monitor, ECG, EDA, EEG and body temperature sensors. They assessed the performance of different features to classify cognitive load on elementary cognitive tasks and found that the ECG median absolute deviation and median heat flux performed best, providing over 80% accuracy [Haapalainen et al., 2010].

In our study, we use sensors for EEG, eye blinks, HR, BVP, EDA and body temperature due to the availability of low-cost and minimally invasive devices and their predictive power.

2.2.3 Interruptibility with Biometric Sensors

Fewer related work used psycho-physiological sensors to assess interruptibility. Kramer gathered EEG data during one hour of US military training and succeeded in classifying interruptibility based on labels gathered retrospectively [Mathan et al., 2007]. Other researchers used measures of an eye tracker to compare mental workload during different hierarchic levels of task boundaries and were able to proof that mental workload dropped at high level task boundaries. This suggests that high level context switches are good moments for interruptions [Bailey and Iqbal, 2008]. Chen et al. conducted a beeper study in which participants solved short tasks with varying difficulty. They measured heart rate variability as indicator for cognitive load and muscle activity trough EMG and calculated interruptibility using linear regression [Chen et al., 2007]. They also developed a mobile phone, which classifies interruptibility into four states combining high / low mental load with high / low movement [Chen and Vertegaal, 2004]. The studies conducted by Bailey et al. [Bailey and Iqbal, 2008] and Chen et al. [Chen et al., 2007] were situated in a controlled laboratory environment and used well defined and relatively simple tasks, such as document editing based on specified comments, typing a given text or solving a word puzzle. Mathan et al [Mathan

et al., 2007] conducted a study during military training with more complex tasks, requiring different cognitive and also physical skills.

Our work differs in that it uses real-world software development tasks requiring a multitude of cognitive skills. In our field study, participants were wearing biometric sensors while working normally in their own offices on their own software development tasks with varying difficulty and context.

2.2.4 Interruption Management

Some related work already used interruptibility indicators in practical applications to avoid interruptions at unsuitable moments. For instance, a decision rule based software delivered unrelated instant messages only at times of context switches that were determined based on mouse movement and window switching. In their study they achieved a five times higher answer rate to messages compared to non-mediated message delivery [Arroyo and Selker, 2011].

Chen and Vertegaal automatically set the ring tone volume of a mobile phone using an interruptibility indicator based on heart variability and muscle activity. In a six person trial they found that participants where satisfied with the chosen notification level in 83% of the cases [Chen et al., 2007].

Only recently, novel designs to handle interruptions from incoming calls on smart phones have been developed adding new actions, such as 'postpone' or 'send a message', to traditional ones, such as 'accept' and 'decline' [Böhmer et al., 2014]. Our work can leverage existing interruption management support by providing an automated interruptibility measure using biometric sensors.

2.2.5 Interruption, Resumption and Edit Lag

There are three time spans commonly used in studies concerning the effects of interruptions: interruption, resumption and edit lag. Interruption lag is the timespan between a notification and the start of the interruption [Altmann and Trafton, 2004]. Resumption lag is the timespan between the end of the interruption and the beginning of the suspended primary task, usually measured by monitoring the first mouse or keyboard action [Adamczyk and Bailey, 2004,

Altmann and Trafton, 2004]. Edit lag is the timespan between the end of the interruption and the first edit and represents a specialized measure of the resumption lag [Parnin and Rugaber, 2011]. It is based on the assumption that interruptions in software development have a large effect and it takes minutes to regather context as opposed to resumption lag which is in the order of seconds [van Solingen et al., 1998].

For immediate interruptions (e.g. a phone call), the interruption lag is very short. Even short interruption lags (8s) are mitigating the disruptiveness of an interruption and can shorten the resumption lag [Trafton et al., 2003]. For negotiated interruptions, the length of the interruption lag can be chosen by the interrupted person [McFarlane, 2002]. It can serve as indicator for interruptibility, following the notion of memory externalization before addressing an interruption [Fogarty et al., 2005b]. The larger the memory load, the less interruptible a person is and the longer the interruption lag is required to find a suitable breakpoint. It has been shown that resumption lag is influenced by the availability of cues [Altmann and Trafton, 2004], the interruption length and demand [Monk et al., 2008], but not by stress, time pressure and flow [Conard and Marsh, 2010].

In our work, we measured interruption lag, the traditional resumption lag and the specialized edit lag to report their lengths and analyze their correlation with interruptibility and mental load before the interruption.

2.3 Study Design

To learn about the interruptibility of software developers and the use of biometric sensors to measure their interruptibility in a real-world context, we conducted two studies, a lab and a field study.

The lab study was a first step to investigate whether biometric sensors can be used to measure interruptibility of software developers working on the same real-world change tasks in a controlled environment.

As a second step, we conducted the field study to investigate how well results from the lab can be transferred to a real-world environment. In both studies,

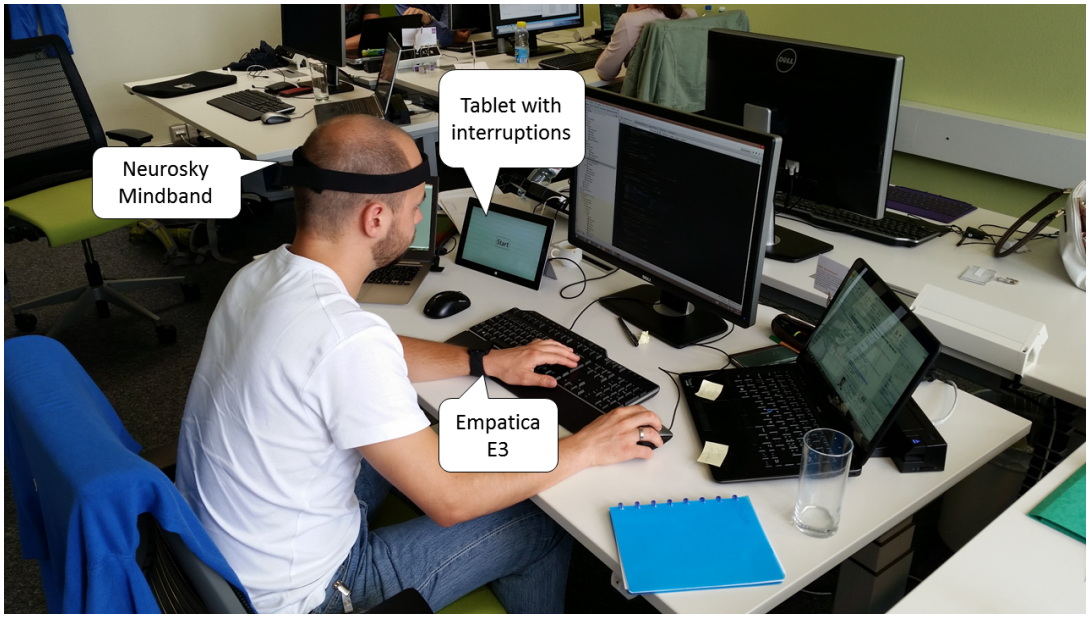


Figure 2.1: Study setup for one participant wearing the headband and wrist band in the field study. The tablet for triggering interruptions is placed next (left) to the participant's main screen.

participants were wearing biometric sensors. Participants were interrupted at random times and asked to perform short arithmetic exercises, as well as rate their interruptibility, their mental load and the disturbance of the interruption. Figure 2.1 illustrates the study setup with one participant from the field study.

2.3.1 Psycho-Physiological Sensors

In both studies, we used two sensor devices: the Neurosky Mindband (http://neurosky.com), a headband to record electroencephalograph (EEG) and eye blink data, and the Empatica E3 wrist band (www.empatica.com) to record electrodermal activity (EDA), skin temperature, blood volume pulse, interbeat interval, and heart rate.

EEG and Eye-Blinks

To measure the electrical brain activity and eye blinks we used the Neurosky Mindband, a low-cost and minimally invasive[1] headband with a one-channel EEG sensor. This headband works with one reference electrode at the ear and two dry electrodes placed on the forehead reading signals mainly from the pre-frontal cortex. The device records the time-varying voltage signal sampled at 512Hz as a raw wave and as a wave filtered for noise. At the same time it records the signal quality that indicates the proper placement of the device. In addition, the headband also records two proprietary measures called Attention and Meditation that are both in the range from 0 to 100, sampled at 1Hz, and meant to indicate mental focus and mental calmness or relaxation respectively.

Skin- and Heart-Related Measures

To record skin- and heart-related measures, we used the wireless Empatica E3 wrist band that integrates an EDA sensor, a photoplethysmograph and a temperature sensor.

The EDA sensor is used to measure skin conductance. It consists of two electrodes that are placed at the ventral area of the wrist. By applying a small current to the skin through these electrodes, skin conductance is measured down to 0.1μS at a frequency of 4Hz. The photoplethysmograph is an optical sensor for measuring blood volume pulse (BVP), which can be used to compute interbeat interval (IBI) and heart rate (HR).

2.3.2 Interruptions

To trigger interruptions during the studies we used a Windows Surface 2 tablet[2] that we placed close to the participant's main monitor. For each interruption, the tablet played a sound and the display changed from a black to a white screen with a "Start" button on it. Participants were instructed to decide for themselves when to address the interruption—a technique for coordinating interruptions

[1] relative to other EEG sensor devices

[2] http://www.microsoft.com/surface/en-us/products/surface-2

called "negotiated interruption" [McFarlane, 2002]. This technique usually works well and was used by others to simulate "normal" interruptions for software developers [Fogarty et al., 2005b].

Using negotiated interruptions allows to measure interruption lag, which is the timespan between the notification and the moment a participant starts to address the interruption [Altmann and Trafton, 2002]. The interruption lag has previously been used as a measure for interruptibility, since researchers found that it corresponds to the time used to externalize the working memory before addressing the pending interruption [Fogarty et al., 2005b].

Interruptions were composed of two parts, a mental arithmetic exercise and a set of questions on the participant's perception of the interruption on five-point Likert scales. For the arithmetic exercise participants were asked to solve a two-digit multiplication. As mental arithmetic exercises generally impose a high working memory load [Lemaire, 1996], they are considered an effective interruption for software developers [Fogarty et al., 2005b]. After participants typed an answer into a text box on the tablet and clicked the "Ok" button, the correct solution was displayed to satisfy participants' need for closure [Kruglanski, 1990].

For the question set part, the tablet application prompted participants to rate their perceived disturbance from 1 (not at all disturbing) to 5 (very disturbing), their interruptibility at the time of the notification from 1 (highly interruptible) to 5 (not at all interruptible), and their mental workload at the time of the notification from 1 (very low) to 5 (very high). After answering these questions, the tablet application displayed a black screen and participants returned to their work.

2.3.3 Lab Study: Participants and Method

For the lab study, all participants worked on the same three real-world code change tasks in the same controlled environment. The study took place in a quiet office room with external distractions and interruptions, except for the ones triggered as part of the study, reduced to a minimum. All participants worked

at the same computer with the same integrated development environment (IDE) setup for the study.

Through personal contacts, we recruited ten graduate students, one female and nine male. All participants had their major in computer science and an average of 4.1 years (standard deviation, in the following denoted with ±, of 3.8) professional and of 10.4 years (±3.2) total development experience. We compensated the participants for their effort with a small chocolate gift.

The study lasted 90 minutes per participant and had three parts, a preparation phase, a 60 minutes programming session and a brief follow-up interview. In the preparation phase, we asked each participant for some demographic information. Then we helped the participant to put on the two sensor devices and had the participant watch a movie of fish swimming in a fish tank for two minutes. The movie was intended to help participants relax and to record a baseline for each participant. We used this later on to normalize the captured data, for instance, to make heart rates comparable among participants with varying resting heart rates.

To familiarize participants with the interruptions, we conducted a few test runs while displaying the fish tank movie. For the main part of the study, each participant was asked to work for 60 minutes in the Eclipse IDE on three code change tasks that we explained to them beforehand. During this programming session participants were frequently interrupted after random time intervals that were between one and eleven minutes long. These time intervals model interruption frequency occurring in reality [González and Mark, 2004]. At the end of the study, we briefly interviewed each participant. In the interview, we asked participants about their perception of the primary tasks, the interruptions with the peripheral arithmetic tasks, as well as more generally about the disruptiveness of interruptions and tool support they would desire for better managing interruptions.

Project and Tasks

We chose the drawing framework JHotDraw[3] for the three code change tasks. JHotDraw is an open source project that has evolved over several years, is well structured and allows for easy testing due to its graphical user interface. The three study tasks were chosen to represent real code change tasks with varying difficulty levels to stress various levels of mental load and various states of interruptibility in a participant.

Adding Circles. The first task is to add a drawable figure, namely a circle, to the Draw application. The task requires to add a button with a provided icon to the toolbar and code to draw the circle. As there is already a feature for drawing an ellipse, code can be reused and the main difficulty is to identify the right places where new code needs to be inserted.

Adding Hexagons. The second task is similar to the first one, but requires to add a hexagon instead of a circle. Knowledge obtained in the first task could be reused, however, drawing a hexagon is more difficult and requires knowledge in geometry. As an optional help, a document with explanations of the geometry of a hexagon was provided.

Adding Text and URL. The third task is to add text and a clickable URL into an existing message dialog. The main task difficulty is to locate the right place in the code for implementing the functionality, and to get familiar with the Java API on message dialogs as well as user interface components.

To validate that the perceived task difficulty varied between tasks, we asked participants to rate them from 1 (very easy) to 5 (very difficult). They rated the first and third task as rather easy (2.4±0.7 and 2±1.4) and the second as rather difficult (3.9±1.0).

[3]http://www.jhotdraw.org/

2.3.4 Field Study: Participants and Method

To learn more about interruptibility and the use of psycho-physiological sensors in the field, we conducted a study with ten professional software developers working on their own tasks and in their real-world office environments. For this study, we visited professional developers in their work places and studied them for two hours each. We did not restrict any external influences, such as interruptions by co-workers or distractions, and we did not limit the work or the participant's work setup, such as the activities they performed during work, the IDEs they used or the office layout.

We recruited ten professional software developers (1 female, 9 male) between their early twenties and late forties from four software development companies of varying size. Participants were recruited through personal contacts and recruiting emails. They had an average of 8.5 years ($\pm$7.5) professional software development experience and an average of 12.7 years ($\pm$6.0) of total development experience. We compensated the participants for their effort with a small chocolate gift.

The field study lasted 2.5 hours per participant and had again three parts, a preparation phase, a main study session of 2 hours and a brief follow-up interview. As in the lab study, we first asked each participant for some demographic information. We then helped them to put on the sensors and had them watch the fish tank movie to help them relax, record a baseline and familiarize participants with the interruptions. For the main study session, participants worked on their usual tasks and at their usual location while being frequently interrupted after random time intervals that, again, were between one and eleven minutes long. For this session participants were told to work as usual during their work day without restriction on their activities, such as checking emails or browsing the web, and to switch tasks as they would normally do. Also, we told all participants and co-workers beforehand to interact during the session as they would usually do during their work day. During the study session, one researcher was present to ensure that everything would run as expected (i.e. that the sensors were recording throughout the whole session) and to observe the time spans needed to recover from the interruptions. At the end of the study, we again conducted a

brief follow-up interview with participants on the perception of the tasks, the disruptiveness of interruptions and desired tool support.

Tasks

During the main study session, participants worked on a variety of tasks, such as the elimination of a performance bottle neck in a web application, the implementation of a user interface component, or the implementation of test cases for a business application. Most participants worked mainly on one primary task during the two hour session and only rarely switched to other small tasks. For these tasks, they used a variety of tools, such as IDEs (e.g. Visual Studio or Eclipse), revision control tools (e.g. SourceTree), web browsers (e.g. Firefox) and email or calendar clients (e.g. MS Outlook).

2.3.5 Data Collection and Analysis

During the course of both studies we collected data from the psycho-physiological sensor devices and data on participants' computer interaction captured through monitoring and observation. In addition, we collected participants' answers to the questions during the interruptions, their demographic information and notes from the brief follow-up interviews. We recorded a total of 30 hours of psycho-physiological data, 10 hours for the 10 lab study participants and 20 hours for the 10 field study participants.

The psycho-physiological data for two lab study participants was not recorded successfully for the entire session. Therefore, we will only present the analysis and results of the 72 valid interruption samples collected from the 8 other lab study participant. The distribution of interruptions is fairly constant per study, with an average of 9 ($\pm$1.6) per person for the 1h lab study and 13.9 ($\pm$2.7) for the 2h field study.

Psycho-Physiological Data

For both studies, we captured the psycho-physiological data with the same sensors and applied the same data cleaning and analysis steps. We discuss these in the following.

EEG and Eye. The raw signal from the EEG sensor is sampled at 512 Hz. We applied a 50 Hz notch filter to remove noise and then split the signal into five commonly used brain wave frequency bands using Matlab's pwelch function: α (8-12Hz), β (12- 30Hz), γ (30-80Hz), δ (0-4Hz), and θ (4-8Hz) [Handy, 2005]. Additionally, we computed fractions of all combinations of frequency bands and the two ratios $\theta/(\alpha+\beta)$ and $\beta/(\alpha+\theta)$ that have previously been shown to carry information on a person's mental activity [Kramer, 1991, Lee and Tan, 2006].

Following a technique suggested by Manilov [Manoilov, 2007], we extracted eye blinks from the raw signal using a Butterworth filter and a peak finding algorithm and calculated the number of eye blinks per time unit. Finally, we used the pre-computed Attention and Meditation signals from Neurosky and extracted the minimum, maximum, mean and standard deviation.

Skin and Heart. The EDA signal generally serves as a measure for arousal and has two main components, the low frequency tonic part that changes over a period of minutes and the higher frequency phasic part that changes within seconds [Schmidt and Walach, 2000]. After filtering noise by applying an exponential smoothing filter, we used a Butterworth filter to split the EDA signal into its tonic and phasic part. From the tonic signal we extracted the skin conductance level (SCL). From the phasic signal we extracted features related to the peaks, in particular the number of peaks per time unit, the mean and the sum of peak amplitudes, and also calculated the area under the curve (AUC). Based on the skin temperature data captured from the integrated thermometer, we extracted the mean temperature and amplitude features, such as the mean and maximum. Based on the captured BVP data we calculated several features, such as the number of peaks per time unit, the mean peak amplitude as well as the heart rate, its mean and variance. Finally, from the IBI, we computed features of heart

rate variability, such as the standard deviation of the signal (SDNN) and the percentage of successive IBIs with a difference greater than 20ms (PNN20) and 50ms (PNN50) [Camm et al., 1996]. All extracted features have previously been linked to various cognitive and emotional states (see Related Work).

Normalization. Since we train our machine learning classifier across participants and psycho-physiological data is very individual, we normalized the data per participant. Therefore, we collected baseline measures during the time each participant watched the relaxing fish tank movie. Normalizing a feature's value for an interruption was then done by subtracting the feature's value calculated using the baseline data from the one calculated using the interruption data.

Interruption, Edit and Resumption Lag

To calculate interruption, edit and resumption lags for participants, we needed to capture five time stamps for each interruption: when the notification for the interruption occurred ($T_{notification}$), when the participant started to address the interruption ($T_{IntStart}$), when the participant finished with the interruption (T_{IntEnd}), when the participant continued to work on her or his primary task using a mouse click or a keyboard action ($T_{FirstInteract}$) and when the participant made the first edit after the interruption ($T_{FirstEdit}$). Interruption lag can then be calculated as $T_{IntStart} - T_{notification}$, resumption lag as $T_{FirstInteract} - T_{IntEnd}$ and edit lag as $T_{FirstEdit} - T_{IntEnd}$. We captured the notification and the interruption start and end with the tablet application. For the lab study, we captured the first interaction and edit with a monitoring software that we installed on the lab study computer and which recorded screen shots and logged mouse and keyboard actions. For the field study we captured the two time stamps by observing participants interactions.

Outcome Measures

To classify interruptibility, both the interruptibility ratings and the interruption lag could serve as ground truth. Although the interruption lag might seem to be a more objective measure, we decided to use the rating in the classification for two

reasons. First, the majority of prior work used the same rating and we wanted to provide a comparable measure, and second, even though interruption lags and interruptibility ratings correlate, they are distributed differently. While ratings are distributed binomially, interruption lags are distributed exponentially. This supports our observation during the studies with some participants addressing notifications quickly, regardless of their interruptibility.

Therefore we believe that the ratings represent the interruptibility more accurately and independent of the participant's interruption handling behavior.

Since we are also interested in examining whether mental load and interruptibility are positively correlated as previous work suggests [Bailey et al., 2001] and whether interruptions during high mental load are more disturbing, we collected participants' ratings of mental load and perceived disturbance as secondary measures. In addition, we collected the interruption, edit and resumption lags as described above to examine findings of prior work on their relation to interruptibility.

2.4 Results

In this section we report on the results of automatically measuring interruptibility using psycho-physiological sensors, the links between interruptibility, mental load and the various lags as well as on the timing and support for interruptions desired by participants.

2.4.1 Measuring Interruptibility

To investigate the use of psycho-physiological sensors for measuring the interruptibility of knowledge workers in a real-world context, we applied machine learning to the collected data in a post-hoc analysis. In particular, we are interested in classifying interruptibility into two states (interruptible or not) as well as a more fine-grained classification using five interruptibility states from highly interruptible to not at all interruptible. For the two state classification we categorized data ratings from 1 (highly interruptible) to 5 (not at all interruptible) into two

groups by labeling data with ratings of 1 to 3 as interruptible and ratings 4 and 5 as not interruptible.

This categorization results in software developers being interruptible in 75% of the samples (83% for the lab and 71% for the field study). For the five state analysis we labeled data with participants' five point ratings of interruptibility. We analyzed the two data sets collected from the lab study (n=72 interruption samples) and the field study (n=139) separately.

Time Window and Classification Algorithm

To examine which time window of psycho-physiological data per interruption works best for the classification, we applied machine learning to several time windows, ranging from ten seconds to three minutes all ending with the notification (see Figure 2.2). Taking into consideration two and five state classification as well as lab and field study results, a time window of ten seconds works generally better than longer ones, except for the five states case of the field study in which longer time windows perform better, probably due to more frequent noise artifacts. The results also show that there are no big differences in accuracy across various time windows. In addition, we examined the use of three different machine learning algorithms: Naïve Bayes, Decision Trees and Support Vector Machine based on Weka's implementations. Overall, Naïve Bayes outperformed the other two algorithms, which is why we will focus on Naïve Bayes in the following.

Validation Methods

For the classification we used a Naïve Bayes implementation of the Weka machine learning framework [Witten et al., 2016]. We applied ten times ten-fold cross-validation, where instances of the participants were distributed randomly across the folds using stratification ('per instance' cross-validation). We chose this approach to investigate the feasibility of using psycho-physiological sensors to predict interruptibility for a development team. This method could be used in a real-world scenario by initially gathering about two hours of data per developer

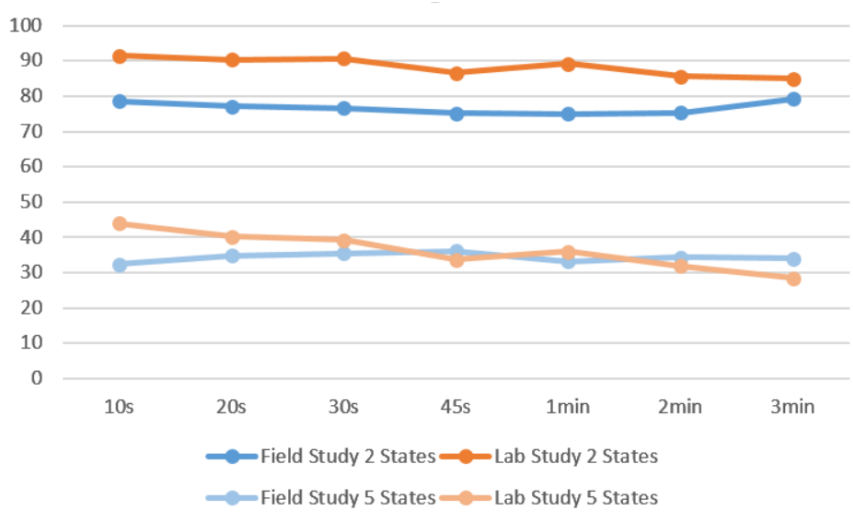


Figure 2.2: Classification accuracies for two and five categories using different time windows and Naïve Bayes.

on the team to train a classifier and then using it to classify interruptibility based on real-time sensor data for developers on the team.

To gather further insights on the generalizability and the sensitivity of the sensors across individuals, we also performed 'per participant' cross-validation. To prevent overfitting, we applied a nested correlation-based feature selection technique (Weka's CfsSubsetEval) that chooses features with high correlation with the class variable but low correlations among each other, using data from the training set of each fold.

Two and Five States Interruptibility

The results of our approach using Naïve Bayes and a ten second time window of psycho-physiological data for each interruption are presented in Table 2.1. For both validation methods, we calculated the accuracy, Cohen's Kappa and the standard deviation of the accuracies obtained during each fold and run. For both studies, we are able to classify the interruptibility at the sample points with high accuracy into two states (91.5% for the lab and 78.6% for the field study) as well as into five states (43.9% for the lab and 32.5%) for 'per instance' cross-validation.

# States	Study	Per Instance CV			Per Participant CV			Majority Classifier
		Accuracy	Cohen's Kappa	Stdev	Accuracy	Cohen's Kappa	Stdev	Accuracy
Two	Lab	91.5%*	0.65	0.7	74.9%	-0.11	36.0	83.3%
	Field	78.6%*	0.44	1.3	69.4%	0.22	19.7	70.7%
Five	Lab	43.9%*	0.18	2.9	37.6%	0.11	21.7	38.9%
	Field	32.5%	0.13	2.1	28.2%	0.07	14.2	32.4%

Table 2.1: Classification results by number of states and study, for per instance and per participant cross-validation (CV), compared to a majority classifier as a baseline value (* indicates that there is a significant difference in accuracy to the majority classifier).

These classifiers perform significantly better than simple majority classifiers for both studies and number of states, except for five states classification in the field. We believe the lack of significance in the latter case is due to more frequent noise artifacts in the field that make it more difficult to distinguish between the finer grained levels.

On the other hand, 'per participant' cross-validation does not result in a significant performance difference compared to a majority classifier and reveals a large variance among participants.

Tables 2.2 and 2.3 show the confusion matrices for the classification per instance, along with F-measures depicting individual class prediction performance. Green cells with bold-faced font indicate correct predictions, orange cells indicate wrong predictions. A darker background color correspond to a larger number of data points. For the classification into five states, the confusion matrix reveals that a great part of the errors is caused by prediction of an adjacent state (cells next to the correct ones), and only few errors are severe (cells far from the correct ones), which is very promising.

Table 2.4 presents the features that were selected at least once by the nested feature selection and for which classifier they were used for in the per instance classification.

The table shows that β, γ and the mean temperature were valuable in all scenarios. EDA related features were only chosen for the lab study data, which might be due to the more frequent occurrence of noise artifacts in the field and the short time windows chosen.

Truth	Prediction: interruptible	Prediction: not interruptible
interruptible	***58.9***	1.1
not interruptible	5	***7***
F-Measure	0.95	0.70

Truth	Prediction: interruptible	Prediction: not interruptible
interruptible	***88.2***	10.8
not interruptible	19	***21***
F-Measure	0.86	0.58

Table 2.2: Confusion matrix for Naïve Bayes classification into two states using per instance cross-validation for the lab study (left) and the field study (right) with individual class accuracies (F-measure). Green cells with bold-faced font indicate correct predictions, orange cells indicate wrong predictions. A darker background color correspond to a larger number of data points.

Truth	Prediction: 1	2	3	4	5
1	***1.2***	1.9	2.8	1.1	0
2	1.5	***9.5***	13.4	0.5	0.1
3	1.9	7.7	***18***	0.4	0
4	1	2	1.2	***0.2***	2.6
5	0.1	0.1	1	1.1	***2.7***
F-Measure	0.19	0.41	0.56	0.04	0.52

Truth	Prediction: 1	2	3	4	5
1	***2.4***	12.4	2.2	1	1
2	3.2	***22***	6.5	2.3	11
3	2.9	12.7	***5***	3.1	11.3
4	1.9	4.3	6.6	***5.3***	6.9
5	0.1	2.3	1.7	0.5	***10.4***
F-Measure	0.16	0.45	0.18	0.28	0.37

Table 2.3: Confusion matrix for Naïve Bayes classification into five states using per instance cross-validation for the lab study (left) and the field study (right) with individual class accuracies (F-measure). Green cells with bold-faced font indicate correct predictions, orange cells indicate wrong predictions. A darker background color correspond to a larger number of data points.

Overall, these results also show that all sensors can provide value for predicting interruptibility.

2.4.2 Interruptibility, Mental Load and Lags

In both studies, there were significant and high positive correlations between the participants' ratings of the perceived disturbance, the mental load and the interruptibility (Pearson's $r > 0.7$, $p<0.0001$). For instance, participants' ratings on interruptibility were highly correlated with ratings on mental load in the lab (r=0.815) and in the field (r=0.702), as well as with ratings on disruptiveness in the lab (r=0.807) and in the field (r=0.741). These high correlations support the often assumed link between mental load and interruptibility, that moments

Sensor	Used Features
EEG	α [2L, 5L], α/γ [2F], α/δ [2F], β [2L, 2F, 5L, 5F], β/α [2L, 2F, 5L], β/γ [2F, 5F], β/δ [2F, 5F], β/θ [2L, 5L], γ [2L, 2F, 5L, 5F], γ/α [2L, 2F, 5L], γ/β [2F, 5F], γ/δ [2F, 5F], γ/θ [2F, 5L], δ [5L], δ/β [2L, 5L], δ/γ [2L, 2F, 5L], δ/θ [2L, 5F], θ [2L], θ/α [2L], θ/δ [2L], $\beta/(\alpha+\theta)$ [2F, 5F], Mean Attention [2F, 5F], Stdev Attention [2F], Min Meditation [2L]
Photoplethys-mograph	Mean Peak Amplitude BVP [2F], Sum Peak Amplitude BVP [2L, 5L], Max Peak Amplitude BVP [5F], Mean HR [2F], IBI PNN20 [2L, 2F], IBI NN50 [2F],
Temperature	Mean Temperature [2L, 2F, 5L, 5F]
EDA	Mean Phasic Peak Amplitude EDA [2L, 5L], Sum Phasic Peak Amplitude EDA [2L, 5L], Phasic Peak Frequency EDA [2L, 5L]

Table 2.4: Most predictive features for Naïve Bayes classification for per instance cross-validation, and their use in the classifiers (2L/2F: lab/field study two states, 5L/5F: lab/field study five states).

of low mental load are suitable for interruptions, and that interruptions during moments of low mental load are perceived less disturbing [Bailey et al., 2001].

Interruptibility was also positively correlated with the interruption lag for the lab study (Pearson's r_{lab}=0.382, p<0.001) and the field study (r_{field}=0.282, p<0.001).

Participants generally took advantage of the so-called negotiated interruption, with an average interruption lag of 30.4 seconds (±7.5) in the lab study, and 44.9 seconds (±6.7) in the field study. Furthermore, 17 of 20 participants commented that they use the interruption lag to finish the current edit in most cases, which overlaps with Fogarty et al.'s finding that participants externalize their working memory before addressing a negotiated interruption [Fogarty et al., 2005b]. These findings—the positive correlation between interruptibility and interruption lag and the evidence for a longer interruption lag corresponding to a higher working

memory load—further support the link between interruptibility and working memory load and thus mental load.

Researchers also found a possible link between resumption/edit lag and interruption lag or interruptibility [Trafton et al., 2003]. We did not find any strong support for this link across studies. There was only a significant correlation between resumption and interruption lag (r=0.275, p=0.001) in the field study and a significant correlation between edit and interruption lag (r=0.251, p=0.04) in the lab study. No other significant correlations at the level of 0.05 (two-sided) were found, including correlations with participants' ratings of interruptibility, mental load or disruptiveness.

2.4.3 Interruption Timing and Support

We used the follow-up interviews to learn more about the cost of interruptions at certain moments and possible tool support. In particular, we asked participants to rate certain situations that we identified in previous literature from 1 (strongly like) to 5 (strongly dislike) and found that, similar to findings of previous studies, participants like interruptions at the end of a task (1.5±0.6) but not in the middle (4±0.7), as well as they dislike them when the mental load is high (5±0) and/or the current task is difficult (4.4±0.9). In situations where participants are stuck and are not making any progress, they feel more mixed about interruptions (2.9±1.2) and several participants mentioned to dislike interruptions in these situations although they stated that interruptions would usually be beneficial for their task and for gaining a different perspective.

When we asked participants about the kind of support they desired for interruptions (open-ended question), they mentioned a tool that displays your interruptibility to co-workers, for instance, by using a lamp (mentioned by 7 participants), and a tool that turns interruptions on or off based on the current mental load (mentioned by 5 participants). In particular, support for in-person interruptions is more needed than for computer based ones since they are generally perceived more disruptive and cannot be ignored. However, participants also commonly mentioned that important interruptions should not be blocked even in situations of extremely high mental load, and that the company culture should

be respected by the tool, e.g., a tool should not prevent interactions that foster team spirit.

2.5 Discussion

The primary focus of the presented work was to investigate the use of psycho-physiological sensors to measure the interruptibility of knowledge workers in a real-world context. The results from our lab and our field study show that using these sensors, we are able to generate machine learning classifiers that can identify a software developer's state of interruptibility—for two as well as five states–with high accuracy. The fact that the results for the lab study are better than the ones for the field study is possibly an indicator for the effect that external influences can have on such sensors, but might also stem from other factors. In particular, we collected a larger amount of data points and also had a higher diversity in the age of the participants in the field study, which could potentially have influenced the observed differences in the prediction performance. The poor performance of 'per participant' cross-validation indicates the high sensitivity of psycho-physiological sensors to individual differences. We assume that much more data is needed to investigate whether it is possible to generalize interruptibility classification using psycho-physiological sensors across different individuals. We believe that our primary choice of 'per instance' cross-validation, for which we train a classifier with data from one team, represents a reasonable trade-off between effort, limitations and value and is applicable in a real-world scenario.

The overall high accuracy for both studies, the task variety and real-world office environment in the field study as well as the use of representative real-world tasks in the lab study, show that these sensors have great potential for measuring interruptibility in a real-world context. As main usage we imagine to display the interruptibility state in real-time via a "traffic light" lamp or IM status, which can potentially help avoiding costly personal interruptions at inopportune moments. Another possibility is to automatically adapt notification settings based on the current interruptibility, where the priority of interruptions has to

be taken into account to not miss important ones. Especially the small time windows of ten seconds required to measure interruptibility with a high accuracy show that these sensors provide the possibility of a real time interruptibility index that can be used for such purposes.

Although our studies were limited to the software development domain, the mobility of these sensors allows for their use in a broad range of domains without being bound to a specific task, computer platform or location and are technically less restrictive than more context-aware sensors, such as a table top sensor or computer interaction monitors (e.g., [Tani and Yamada, 2013, Fogarty et al., 2005b]). We believe that our results are therefore an encouraging step in the use of such sensors in a real-world work context and warrant further research on applying these sensors to a broader range of knowledge workers.

The results show that measures from the EEG, Photoplethysmograph, temperature and EDA sensors can all provide valuable information for classifying interruptibility. While EEG measures were selected for all classifiers, the EEG headband might be too obtrusive for long-term use. In future work, we plan to examine the use of subsets of sensors over longer periods of time to achieve a high accuracy and usability.

In addition to providing evidence for the benefits of using these sensors to measure interruptibility, the results of our study also confirm previous results on the correlation between mental load and interruptibility [Bailey et al., 2001, Iqbal and Bailey, 2005], the link between interruption lag and interruptibility in setups with negotiated interruptions [Fogarty et al., 2005b], and the findings that interruptions at task-boundaries are less disruptive [Bailey and Iqbal, 2008].

2.6 Conclusion

In this work we investigated the use of psycho-physiological sensors to automatically classify the interruptibility of knowledge workers in a real-world context. We conducted two studies, a lab and a field study, in which we captured the psycho-physiological data of twenty participants for a total of 30 hours and interrupted them at random times. Using a Naïve Bayes classifier, we are able to

predict the interruptibility of participants with high accuracy, improving significantly upon a majority classifier. Our results also confirm previous findings on the positive correlation between interruptibility and mental load, which further supports the use of psycho-physiological sensors that have already been shown to indicate states of mental load in other studies.

For future work, we aim at designing and prototyping tool support leveraging the predictive power of psycho-physiological sensors to help knowledge workers with their management of interruptions. Based on the interviews of our study, such support is highly desired for direct interruptions from co-workers, but also for computer-based interruptions, such as e-mails and instant messages.

2.7 Acknowledgments

The authors would like to thank all study participants. This work was funded in part by SNF and an ABB research grant.

3

Sensing Interruptibility in the Office: A Field Study on the Use of Biometric and Computer Interaction Sensors

Manuela Züger, Sebastian C. Müller, André N. Meyer, Thomas Fritz
Published at the 2018 CHI Conference on Human Factors in Computing Systems
Contribution: Study design, data collection, data analysis, and paper writing

Abstract

Knowledge workers experience many interruptions during their work day. Especially when they happen at inopportune moments, interruptions can incur high costs, cause time loss and frustration. Knowing a person's interruptibility allows optimizing the timing of interruptions and minimize disruption. Recent

advances in technology provide the opportunity to collect a wide variety of data on knowledge workers to predict interruptibility. While prior work predominantly examined interruptibility based on a single data type and in short lab studies, we conducted a two-week field study with 13 professional software developers to investigate a variety of computer interaction, heart-, sleep-, and physical activity-related data. Our analysis shows that computer interaction data is more accurate in predicting interruptibility at the computer than biometric data (74.8% vs. 68.3% accuracy), and that combining both yields the best results (75.7% accuracy). We discuss our findings and their practical applicability also in light of collected qualitative data.

3.1 Introduction

In today's collaborative work environments, knowledge workers are constantly facing interruptions, such as instant message alerts, emails or a co-worker asking a question in person [González and Mark, 2004, Chong and Siino, 2006, Iqbal and Horvitz, 2007]. Many of these interruptions are necessary to share knowledge and resolve problems quickly [Isaacs et al., 1997]. Yet, the timing of the interruption has a big impact on its disruptiveness [Adamczyk and Bailey, 2004, Bailey and Konstan, 2006]. Several studies have demonstrated the negative effects that interruptions can have when they happen at inopportune moments, e.g. when a person is highly focused, ranging from a higher error rate and a lower overall performance to more stress and frustration [Bailey et al., 2001, Czerwinski et al., 2000, Mark et al., 2008]. To optimize the timing of interruptions and reduce the disruptiveness and negative effects, researchers have looked into measuring a person's interruptibility—the availability for interruptions. Such an interruptibility measure could then be used to postpone computer-based interruptions to more opportune moments [Iqbal and Bailey, 2008], or to provide awareness to co-workers and prevent in-person interruptions at inopportune moments [Züger et al., 2017].

Prior research on measuring interruptibility can roughly be categorized by the kinds of sensors examined: computer interaction or biometric (*aka.* psycho-

physiological) sensors. Studies investigating computer interaction use features such as keyboard/mouse input or application usage to find suitable moments for interruptions [Fogarty et al., 2005b, Iqbal and Bailey, 2008]. Studies on biometric sensors are based on the assumption that physiological features, such as heart rate, pupil dilation or brain activity, can be linked to the user's cognitive states and task engagement and thus be used to determine interruptibility [Chen and Vertegaal, 2004, Bailey and Iqbal, 2008, Züger and Fritz, 2015]. While study results have demonstrated the potential of features from both sensor types to determine a person's interruptibility, the studies were predominantly conducted on small and controlled tasks over short periods of time (less than three hours) and mostly limited to either computer interaction or biometric sensors.

In the presented research, we build upon and extend previous work by investigating the use of computer interaction and biometric sensors to determine a person's interruptibility at office work-places over a two-week period. Especially since computer interaction sensors are limited to a specific kind of interaction and work during the day and biometric sensors are more physically invasive and more sensitive to noise (e.g. movement artifacts), we are interested in examining the accuracy and feasibility of features of either one or a combination of both sensor types in the field and over a longer period of time. We conducted a two-week field study with 13 professional software developers from three companies, enabling us to study a homogeneous group with similar work patterns, including a variety of activities of which many are performed on the computer [Storey et al., 2017, Bacchelli and Bird, 2013, Vasilescu et al., 2016, González and Mark, 2004]. We collected biometric data from several sensors including heart rate, physical activity and sleep measurements, as well as computer interaction data including mouse and keyboard interaction, the active application window, and time and calendar information. In addition, we collected interruptibility ratings through experience sampling using a pop-up displayed on the computer, that we then used as ground truth for predicting a participant's interruptibility.

With the study at hand, we aim to build a classifier that predicts a software developer's interruptibility accurately in the field. Therefore, we first examine the optimal *time window* to extract features from the continuous biometric and

computer interaction data. Second, we examine the best combination of *sensors and features* using machine learning techniques and how these quantitative results align with the participants' *subjective perceptions* based on qualitative survey and interview data. Finally, we examine whether it is possible to create a *general classifier* rather than one per individual for predicting interruptibility with high accuracy for new people.

In our analysis we found that: (a) the optimal time windows vary per feature (e.g., 10-20min for user input and 2-3min for heart-related data); (b) computer interaction sensors had more predictive power than biometric sensors (74.8% accuracy compared to 68.3% on average), while a combination of both was most accurate (75.7%); (c) participants' perceptions overlap with quantitatively identified feature importance; and that (d) a general classifier can achieve a high accuracy (69.8%), yet a classifier trained for a single individual can outperform the general one even with few data points. Our main contributions are an analysis of predicting software developers' interruptibility in the field, and a comparison of the predictive power of various biometric and computer interaction features.

3.2 Related Work

Related work in the area primarily focuses on studies on interruptions, in particular their effects and factors influencing their disruptiveness, and on approaches to measure interruptibility.

3.2.1 Interruptions at the Workplace

Several observational studies showed that a typical work day of knowledge workers is highly fragmented. On average, they switch activities every 2-3 minutes and get interrupted 13 times a day, e.g. through personal visits, emails or phone calls [González and Mark, 2004]. Solingen et al. found that people spend 15-20 minutes per interruption and a total amount of 15-20% of their time handling interruptions [van Solingen et al., 1998]. Sykes reported that the longest

interruptions are personal visits from colleagues (ranging from 24 minutes up to 4 hours) [Sykes, 2011].

Many interruptions are necessary in a collaborative work space, and often a short interruption can help a co-worker to solve a problem quickly and make progress on a task [van Solingen et al., 1998]. However, interruptions can also have multiple negative effects, such as long resumption lags and an increase in errors and frustration (e.g. [Bailey et al., 2001, Czerwinski et al., 2000]). Often, knowledge workers do not even go back to their suspended task directly after an interruption [Mark et al., 2005], or compensate for interruptions by working faster which leads to more stress and frustration [Mark et al., 2008].

Not all interruptions are equally disruptive. Studies found the interruption moment, duration and frequency as well as the difficulty of the interrupting task and its relevance to current work to be important factors in the disruptiveness of interruptions [Bailey and Iqbal, 2008, Czerwinski et al., 2000, Monk et al., 2008, Cades et al., 2007, Gillie and Broadbent, 1989]. Borst et al. developed a disruptiveness model of interruptions and found that the memory required for the interrupted and interrupting task is an important factor, explaining why interruptions are less costly at breakpoints and times of low mental work load compared to moments in the middle of tasks and during high mental workload [Borst et al., 2015]. With our research we contribute an analysis of automatic and continuous measures of interruptibility in the field that can be used to find opportune moments for interruptions and reduce their disruptiveness.

3.2.2 Finding Opportune Moments for Interruptions

There are primarily two ways to optimize the moment of interruptions: deferring interruptions to task boundaries or continuously measuring interruptibility even during tasks. Since working memory is usually low at task boundaries, the defer-to-boundary policy aims at determining these natural breakpoints during work and delaying interruptions, such as email notifications, to these more opportune moments [Iqbal et al., 2004a, Bailey and Konstan, 2006]. Another type of approaches aims at predicting interruptibility continuously [Fogarty et al., 2005b, Züger and Fritz, 2015]. These approaches are particularly useful to

reduce in-person interruptions at inopportune moments by indicating the current interruptibility state to potential interrupters [Züger et al., 2017], but can also be used to postpone computer-based interruptions from moments of low to high interruptibility.

Approaches to continuously measure interruptibility can broadly be categorized by the types of sensors used: biometric, computer interaction, or context sensors. Biometric sensors can be used to measure the body's activities and responses to external stimuli. Various studies have shown that biometric data such as heart rate (HR), heart rate variability (HRV), electro-dermal activity (EDA), eye tracking, skin temperature or electroencephalography (EEG) can be used to assess mental effort and cognitive load [Wilson, 2002, Richter et al., 1998, Haapalainen et al., 2010, Chen et al., 2007], task difficulty [Veltman and Gaillard, 1998, Fritz et al., 2014], emotions [Maaoui et al., 2010, Haag et al., 2004, Müller and Fritz, 2015], or stress [Healey and Picard, 2005, Sano and Picard, 2013, Wijsman et al., 2011]. A few researchers have also investigated whether such measurements can be used to measure interruptibility. Mathan et al. used an EEG device to compute interruptibility during military training [Mathan et al., 2007]. Goyal and Fussell used EDA data to find opportune moments for interruptions in a lab study [Goyal and Fussell, 2017]. Bailey and Iqbal as well as Katidioti et al. used measures of pupil dilation to find suitable moments for interruptions in lab studies [Bailey and Iqbal, 2008, Katidioti et al., 2016]. In a short lab and field study with software developers, a combination of HR, HRV, EDA, and EEG sensors has been used to predict interruptibility [Züger and Fritz, 2015]. Furthermore, accelerometer data has been used in several studies to detect physical activity and to show that interruptions are better delivered during moments recognized as activity transitions, e.g. when walking to another location [Ho and Intille, 2005, Fisher and Simmons, 2011, Komuro et al., 2017]. A further and not yet fully studied factor of interruptibility is sleep, which has been shown to have a big impact on productivity and mood [Rosekind et al., 2010, Vidaček et al., 1986, Mark et al., 2016a].

Computer interaction sensors measure a user's interaction with task artifacts on the computer. They mainly consist of mouse, keyboard, and application usage

data. Some studies went a step further to get more context from other sources such as audio and video recordings, calendar or network connection data. As an example, Fogarty et al. collected a total of 475 interruptibility ratings and IDE interaction data from 20 participants to measure interruptibility during software development tasks [Fogarty et al., 2005b]. Other researchers identified breakpoints using computer interaction sensor features such as the frequency of window switches in studies ranging from a few hours [Tanaka and Fujita, 2011, Iqbal and Bailey, 2008] to 2 weeks [Nair et al., 2005]. Kapoor and Horvitz developed BusyBody, an approach that calculates interruptibility using a rich set of computer interaction and contextual features from user input, calendar, time and wireless signal data [Horvitz et al., 2004, Kapoor and Horvitz, 2007, Kapoor and Horvitz, 2008]. Horvitz et al. built a query-able service to predict a user's presence and availability from user activity and proximity from multiple devices, calendar and time information [Horvitz et al., 2002]. Another body of research focused on indicating interruptibility or availability in messaging clients or physical indicator lights, and used computer or device interaction, location, speech, calendar, time, presence or network data, or a combination thereof as underlying sensing technique [Züger et al., 2017, Begole et al., 2004, Lai et al., 2003, Fogarty et al., 2004].

In our study, we extend upon prior work by using a combination of biometric and computer interaction sensors in the field. We used two biometric sensors (a Fitbit Charge 2 and a Polar H7) to measure HR, HRV, physical activity and sleep and to capture a wide range of biometric data with little invasiveness, compared to e.g. EEG and eye tracking devices, which are more difficult to use in the field. For computer interaction, we recorded the user input (keystrokes and mouse interactions), application usage, and calendar data. To our knowledge this is the first study using this combination of sensors to investigate the continuous measurement of interruptibility in the field and for a longer period of time, in particular its accuracy, feasibility and the predictive power of various types of data.

3.3 Study Design

To study the prediction of interruptibility in the field, we conducted a two-week field study with 13 professional software developers. For this study, we gathered a rich set of data, including a variety of biometric and computer interaction data as well as interruptibility ratings and qualitative data.

Participants. We recruited 14 software developers through professional and personal contacts from one large-sized and two medium sized companies in the software industry. We focused on software developers as one community of knowledge workers, to ensure our participants have similar work patterns including a wide variety of activities and extensive computer use to support both individual and collaborative tasks [Storey et al., 2017, Bacchelli and Bird, 2013, Vasilescu et al., 2016, González and Mark, 2004]. We discarded the data of one participant due to a technical issue with the Polar sensor that led to no recordings from this sensor and thus an incomplete and incomparable dataset for this participant. Of the remaining 13 participants, 1 was female and 12 were male. At the time of the study, participants had an average age of 32.4 years (standard deviation, in the following denoted with ±, of 6.2), an average professional experience of 6.5 years (± 6.2) and total experience in software development of 11.8 years (± 6.6). Most participants were individual contributors (6), and the others had job roles such as architects (3), executives (1), lead (2) and other (1). We compensated the participants for their effort with a small chocolate gift.

Procedure. At the beginning of the study, we explained the purpose and process of the study, and handed out, set up and introduced the two biometric sensors (Fitbit Charge 2 and Polar H7). We asked the participants to wear the Polar sensor during work hours, and the Fitbit sensor as much as possible including work and free time as well as nights, except when they did not feel comfortable wearing it or when swimming, showering or charging the device. The participants synced the data every one or two days. In addition, we installed a **monitoring tool** to collect computer interaction data. In case a participant worked on several computers, we installed the monitoring tool on all of them to collect a complete data set. We further automated the synchronization of

the time for all devices (computers and sensors) participants used for the study period.

For the following two weeks (some participants also continued the study for a few more days), we asked participants to follow the same procedure every work day. We asked them to wear the biometric sensors, to rate their interruptibility when prompted by a pop-up on their computer with an experience sampling technique, and to fill out a short daily diary survey regarding their perception of the work day in the evening.

At the end of the study, we collected the sensors and data, conducted interviews on our participants' perception on interruptibility, and asked them to fill out our end survey with demographic questions. In the remainder of this section, each part of the study procedure is explained in detail.

Biometric Sensors. Based on prior research as well as invasiveness, we chose to use two biometric sensors for our field study: the Polar H7 for recording HR and HRV data, which both have been linked to stress and cognitive load by previous research [Acharya et al., 2006, Haapalainen et al., 2010], and the Fitbit Charge 2 for recording HR (sampled every 10s), physical activity (sampled every 1min), and sleep (duration and quality metrics), which have been linked to interruptibility [Ho and Intille, 2005, Fisher and Simmons, 2011, Komuro et al., 2017] and productivity [Rosekind et al., 2010, Vidaček et al., 1986].

The Polar H7 [Electro, 2017] is a chest strap recording heart beats and interbeat-intervals, using an electrocardiograph (ECG) based sensor technique with medical grade accuracy [Wang et al., 2017]. The Polar's minimally invasive form-factor and long battery life make it feasible to be used in a field study. Since the sensor has no built in memory, we extended our monitoring tool with the capability to receive the measurements of the device via bluetooth, which limits the data collection with this sensor to the time spent within bluetooth range of the computer.

The Fitbit Charge 2 [Fitbit, 2017] is one of the most accurate wrist-worn activity trackers [Guo et al., 2013]. While the Fitbit's coarser sampling granularity does not allow measuring HRV [Wang et al., 2017] and tends to overestimate sleep duration, it has a high intra-device reliability [Montgomery-Downs et al.,

2012] and can be worn constantly (except for charging, showering and swimming) thanks to the minimally invasive form-factor and the built-in memory. The Fitbit data was synced to Fitbit servers via bluetooth using the official smart phone or computer application, and then automatically downloaded by our monitoring tool. For this purpose, the participants granted our monitoring tool access to the Fitbit account during the study.

Monitoring Tool. To collect computer interaction data, we used our own monitoring tool for the Windows operating system that tracks a participant's mouse and keyboard interactions, the active window, and calendar information. For the mouse, the clicks (coordinates and button), the movement (coordinates and moved distance in pixels), and the scrolling (coordinates and scrolled distance in pixels) are tracked along with the corresponding timestamp. For the keyboard, we recorded the keystroke type (normal, navigating or delete key) along with the corresponding timestamp. We did not record specific keystrokes for privacy reasons. For the active window, we recorded the name of the active process and the window title, along with the timestamp at which the user switched to the window. For calendar data, the tool used the Microsoft Graph API of the Office 365 Suite [Microsoft, 2017] and recorded start time, duration and subject of meetings.

Interruptibility Ratings. To collect the ground truth for the interruptibility classification, we prompted our participants with an experience sampling technique using a pop-up that was displayed on the computer. The prompts asked participants to rate their current interruptibility on a 7-point Likert scale and were displayed in random intervals between 10 and 40 minutes. We chose this time interval as a trade-off between annoyance and invasiveness while also collecting enough samples to apply machine learning. This decision was based on our experience from a pilot study with 8 software developers during 7 work days and from testing the final study procedure ourselves for several days. In the pilot study, we further observed that some participants tended to avoid the extreme or intermediate parts of the scale. Therefore we extended the original 5-point Likert scale (which has predominantly been used in related work [Fogarty et al., 2005b, Tanaka and Fujita, 2011]) to a 7-point Likert scale to obtain a

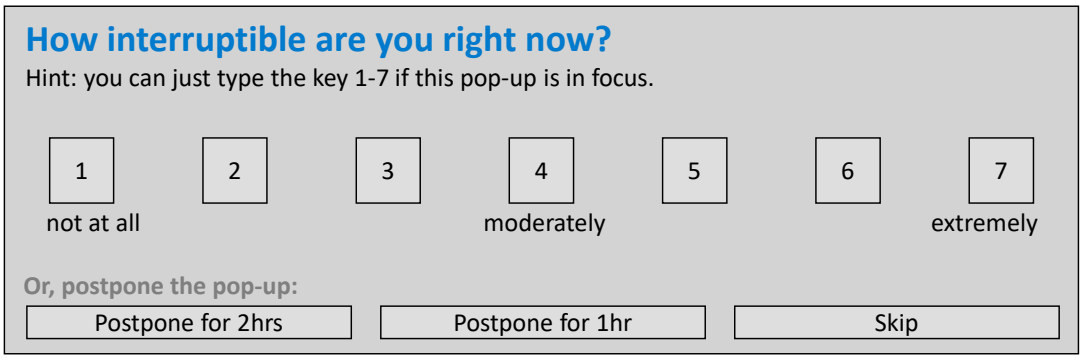


Figure 3.1: Screenshot of the interruptibility rating pop-up

higher variety of ratings. The pop-up prompts were displayed in the bottom right corner of the main screen and were directly integrated into the monitoring tool (see Figure 3.1). With just one click, the prompt could be answered, skipped or postponed for the next one or two hours, preventing false answers caused by annoyance. For participants that used multiple computers simultaneously, we disabled, if desired, the prompts on all except the main computer to prevent fatigue from too many and frequent prompts. Our participants had the possibility to correct a rating by sending an email which occurred twice throughout the course of the study.

Questionnaire and Interviews. We collected qualitative data to gain insights on participants' perceptions of interruptibility and related factors, complementary to the quantitative data. At the end of each work day, the participants answered the same short **diary questionnaire** containing items regarding their work day. Each question was rated on a 7-point Likert scale and included productivity, sleepiness, challenge, engagement, arousal, valence, stress, interruption frequency, and daily interruptibility. As an example, we asked the participants: *Compared to an average work day, how stressed were you today?* We included these items to analyze their relation to interruptibility and chose them based on literature and their potential impact on interruptibility (e.g. [Rosekind et al., 2010, Bailey and Iqbal, 2008, Müller and Fritz, 2015, Mark et al., 2008]). At the end of the study period, we further conducted **interviews** to ask open

	Total	per Participant
Polar data	808 hours	62 hours (± 12)
Fitbit data	5532 hours	426 hours (± 76)
Fitbit sleep data	197 nights	15 nights (± 4)
Computer monitoring data	3552 hours	273 hours (± 143)
Calendar entries	746 meetings	57 meetings (± 37)
Interruptibility ratings	2515 samples	193 samples (± 88)
Interviews	525 minutes	40 minutes (± 8)
Diary survey	151 responses	12 responses (± 1)

Table 3.1: Collected data

questions about factors that influence participants' interruptibility, and about their experience with the biometric sensors and the monitoring tool. The study concluded with an **end questionnaire** to collect demographic data.

3.4 Data Collection and Preprocessing

In our two-week field study, we collected a rich set of quantitative and qualitative data (see details in Table 3.1). Prior to the main analysis of the data, we performed multiple preprocessing steps that are summarized in the remainder of this section. Our preprocessing and analysis scripts along with more detailed explanations and information are available online[1].

Basic Preprocessing. Before analyzing the computer interaction data, we anonymized the data by replacing identifying text fragments with placeholders. In particular, the raw window titles could potentially include names or email addresses, which we replaced with placeholders such as *<name1>* or *<email2>*. We further merged the computer interaction data for participants that worked on several computers in parallel, mostly by adding all data points into one common database. For two participants that used Remote Desktop Connection to switch between computers, we further had to delete entries representing the Remote

[1]http://dx.doi.org/10.5281/zenodo.1118965

Desktop window, and only used the user input from the main machine to prevent duplicates.

Feature Extraction. A first step towards building a reliable interruptibility classifier is to extract meaningful features of the raw data. We extracted features that have previously been linked to cognitive states such as cognitive load, stress or emotions, and also interruptibility. Table 3.2 provides an overview of all 85 features that we extracted along with the corresponding references where they have been defined or used previously.

From the computer interaction data, we extracted user input features, in particular frequency and duration measures of keystroke and mouse events that capture if a person is actively producing content or being idle, e.g. thinking, reading or away from the computer. We further extracted application window features that capture window switching events and time spent in specific activity categories. We define an application window as a unique combination of the process name and window title. An application window switch can therefore refer to a switch between two different applications as well as, for example, a switch between two different tabs in a web browser. We obtained activity categories from the window switching events by mapping window and process names to a general activity category such as *Coding*, *Reading or Writing Documents*, or *Email or Planning* (for all categories see Table 3.2). We used common categories typical for software developers that had previously been identified by Meyer et al. [Meyer et al., 2014]. We mapped the data semi-automatically in two stages. First, an automatic algorithm developed by Meyer et al. mapped obvious programs and activities, such as *Microsoft Visual Studio* belonging to the activity category *Coding* [Meyer et al., 2017a]. In a second step, one author manually mapped the remaining entries using the window titles that provided valuable contextual information, e.g. to distinguish between *Work Related Browsing* and *Work Unrelated Browsing*. We further extracted features related to focus duration and activity / category switching frequency inspired by Sarkar and Parnin who used these features to predict mental fatigue of software developers [Sarkar and Parnin, 2017]. From the calendar entries we extracted features indicating whether

the person had scheduled a meeting for the recent past or future. Finally, to capture data related to the circadian rhythm we extracted time related features, e.g. the hour arrived at work based on the first interaction with the computer per day.

Feature Group	Importance	Features
User Input	**29.6%**	*Sensor: Computer Monitoring*
Keystrokes	11.3%	Number of all (2min, 10min) / normal (20min) / navigation (20min) / delete (20min) keystrokes per min, percentage of time spent typing (10min)
Mouse Clicks	8.2%	Number of all (10min) / left (10min) / middle (45min) / right (30min) / other (20s, 45min) mouse clicks per min, percentage of time spent clicking (10min)
Mouse Scrolls	3.2%	Scrolled distance per min (30min), percentage of time spent scrolling (30min)
Mouse Moves	4.2%	Moved distance per min (20min), percentage of time spent mouse moving (10min)
Keystrokes & Mouse	2.6%	Percentage of time being idle (10min)
		[Iqbal and Bailey, 2007, Shrot et al., 2014, Arroyo and Selker, 2011, Horvitz and Apacible, 2003, Fogarty et al., 2004, Züger et al., 2017, Sarkar and Parnin, 2017, Kapoor and Horvitz, 2007, Kapoor and Horvitz, 2008, Horvitz et al., 2004, Horvitz et al., 2002]
Application Window	**44.6%**	*Sensor: Computer Monitoring*
Activity Category	30.4%	Percentage of time spent in the following activity categories and sub categories: Software development (2min) (coding (3min, 10min), debugging (5min), version control (10min), reviewing (2min)), communicating (3h) (email (1h), instant message (2min, 2h)), reading or editing documents (1min, 20min), web browsing (30s, 20min) (work related browsing (10min), work unrelated browsing (3h)), work unrelated activities (10s) (work unrelated browsing (10s), work unrelated apps (1min, 3h)), planning (10s, 20min, 45min, 3h), navigating and other (45min)
Focus Duration	5.9%	Max. time in one application window (20min) / category (10s, 20min)
Activity Switches	8.2%	Number of application window (20min) / category (10s, 5min, 20min) switches per min, number of distinct categories (10s, 20min)
		[Iqbal and Bailey, 2007, Nair et al., 2005, Mirza et al., 2011, Arroyo and Selker, 2011, Kapoor and Horvitz, 2007, Kapoor and Horvitz, 2008, Horvitz et al., 2004, Horvitz et al., 2002]
Calendar	**2.8%**	*Sensor: Computer Monitoring*
Past Meetings	1.8%	Number of past meetings per hour (3h), percentage of time spent in meetings (7.5min, 3h), meeting now (boolean)
Upcoming Meetings	1.0%	Number of upcoming meetings per hour (1min, 45min), percentage of time planned in meetings (30s, 1h)
		[Stern et al., 2011, Horvitz and Apacible, 2003, Fogarty et al., 2004, Züger et al., 2017, Kapoor and Horvitz, 2007, Kapoor and Horvitz, 2008, Horvitz et al., 2004, Horvitz et al., 2002]
Heart	**14.2%**	*Sensors: Polar and Fitbit*
HR	9.8%	Polar HR mean (20s, 3min) / std. dev. (45s), Fitbit HR mean (20s) / std. dev. (10min), Fitbit resting HR, Fitbit percentage of time spent in HR zones (45min)
HRV	4.4%	Polar SDNN (3min), Polar RMSSD (3min), Polar pNN50 (2min)
		[Wilson, 2002, Haapalainen et al., 2010, Züger and Fritz, 2015, Mulder, 1992, Chen et al., 2007, Healey and Picard, 2005, Acharya et al., 2006, Xhyheri et al., 2012, Karvonen and Vuorimaa, 1988, Association, 2016, Haag et al., 2004]
Movement	**2.3%**	*Sensor: Fitbit*
Steps	2.3%	Number of steps per min (2min), percentage of time spent walking (3min)
		[Ho and Intille, 2005, Fisher and Simmons, 2011, Komuro et al., 2017]
Circadian Rhythm	**6.5%**	*Sensors: Computer Monitoring and Fitbit*
Time	2.1%	Hour of day, day of week, hour arrived at work
Sleep	4.4%	Duration, sleep efficiency, hour of midpoint of sleep, hour of wakeup, number and minutes being awake / restless
		[Visuri et al., 2017, Mark et al., 2014, Pilcher et al., 1997, Rosekind et al., 2010, Vidaček et al., 1986, Mark et al., 2016a, Kapoor and Horvitz, 2007, Kapoor and Horvitz, 2008, Horvitz et al., 2004, Horvitz et al., 2002]

Table 3.2: Features analyzed in our study and grouped by sensor together with the feature's importance for the interruptibility classifier, the used time window per feature (colored and in brackets), and references to prior related work on these features.

From the biometric data we extracted HR and HRV related features from both the Polar and the Fitbit sensors by taking advantage of the higher accuracy of the Polar and the larger amount of data available from the Fitbit. For HRV, we used three standardized metrics: the standard deviation of the successive differences of heart beats (SDNN), the root mean square of the successive differences (RMSSD) and the proportion of pairs of successive intervals that differ by more than 50 ms (pNN50) [Xhyheri et al., 2012]. To calculate the heart rate zones, we used the Karvonen method, using the mean of the daily resting heart rates measured by the Fitbit Charge 2 throughout the whole study period and the age the participants reported [Karvonen and Vuorimaa, 1988]. We use heart rate zones as suggested by the American Heart Association and used by Fitbit: up to 49% of the maximum heart rate is regarded as being out of zone, 50% to 69% is labeled with low activity, 70% to 84% high activity and 85% and more is peak activity [Association, 2016]. Steps and sleep measurements were extracted as indicated in Table 3.2.

Outcome Measure. As outcome measure we used the interruptibility ratings collected with experience sampling. Figure 3.2 shows that prompts were answered throughout the whole work day, though less often early in the morning, at lunch and in the evening and that most prompts were answered quickly (50% within 8s, and 83% within 15 minutes). To predict if a person is interruptible, we reduced the 7-point Likert scale to two states (splitting at 123 | 4567), similarly to previous studies predicting interruptibility based on experience sampling ratings, which split a 5-point Likert scale between 2 and 3, counting the middle rating to the interruptible samples [Fogarty et al., 2005b, Züger and Fritz, 2015]. For our more fine-grained analysis, we used the full 7-point Likert scale and additionally split it into three states (splitting at 12 | 345 | 67). As one participant never used a rating of 1 or 2 and thus had a highly imbalanced dataset using this splitting method (for two states: 91.4% being interruptible - 8.6% being non-interruptible), we accommodated for the imbalance by using a different splitting mechanism (1234 | 567 and 1234 | 5 | 67) for this participant.

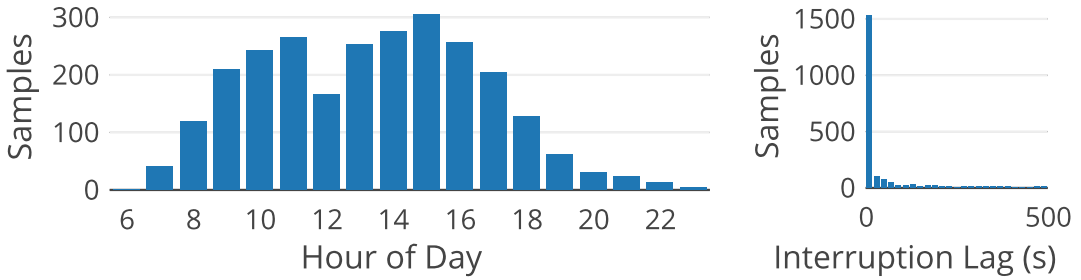


Figure 3.2: Distribution of self-reports and interruption lags (truncated after 500s for better readability).

Machine Learning Tuning. We used scikit-learn [Pedregosa et al., 2011], a widely used machine learning library for Python, to predict interruptibility from biometric and computer interaction data. We evaluated several classifiers by applying them to our feature set and testing different parameter values. A random forest classifier (500 estimators, no prior feature selection) outperformed all other approaches, including a gradient boosting classifier (500 estimators, max. depth=3, no prior feature selection), support vector machine (kernel=RBF, C=1, gamma=0.03, selected 30 best features prior to classification), neural network (solver=LBFGS, alpha=0.0001, hidden layers=100, no prior feature selection) and Naïve Bayes classifier (selected 30 best features prior to classification) [Sammut and Webb, 2011]. Therefore, for the remainder of this paper, we will present results obtained with a random forest classifier. A random forest classifier is an ensemble learning method that creates a multitude of decision tree classifiers and aggregates their predictions with a voting mechanism [Breiman, 2001, Liaw et al., 2002]. It is noteworthy that this classifier does not require preselecting features, and can deal with a large feature space that also contains correlated features. In all our machine learning experiments, we first imputed missing values by replacing them with the mean, and normalized the features to comparable scales using a *StandardScaler*. These are common initial steps in a machine learning pipeline and a requirement for many classifiers to work properly [Pedregosa et al., 2011].

3.5 Analysis and Results

To examine whether we can use the collected sensor data to accurately predict interruptibility in the field and which combination of computer interaction and biometric features achieves the highest accuracy, we applied machine learning to our preprocessed features using the self-reports as the outcome measure. In the following, we first examine which time windows to use for each extracted feature, followed by an analysis and findings of the best features and combinations thereof. To complement the quantitative results, we further analyze how participants' perceptions of their interruptibility overlap with our findings. Finally, we investigate how well a general classifier of interruptibility can be used across participants in the field compared to an individually trained classifier and examine whether the features can also be used to predict interruptibility on a more fine-grained level.

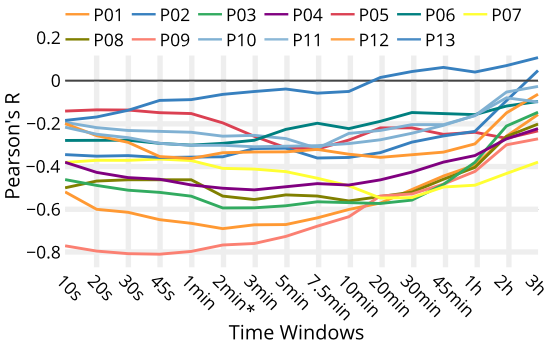
3.5.1 Time Windows

As a first step in determining a classifier for interruptibility, we have to decide on the time windows that are being used for each of the extracted features so that we can transform the continuous data streams of each feature into discrete variables. The time window can have an impact on the classifier as previous research has shown [Vorburger et al., 2011, Züger and Fritz, 2015]. While previous researchers have used a variety of time windows for predicting interruptibility, predominantly between 1s and 5mins [Hudson et al., 2003, Fogarty et al., 2005b, Züger and Fritz, 2015], there is no general guideline on which time windows to use for which feature. In our analysis, we take advantage of the longitudinal nature of our study and the variety of features examined and analyze which time windows are optimal to predict interruptibility. In particular, we analyze an extensive set of time windows ranging from 10 seconds all the way to 3 hours: *10s, 20s, 30s, 45s, 1min, 2min, 3min, 5min, 7.5min, 10min, 20min, 30min, 45min, 1h, 2h, 3h.* We put a focus on shorter time windows due to their use in prior studies, but we also include longer time windows that have not been examined in earlier studies, especially due to the short and controlled nature of

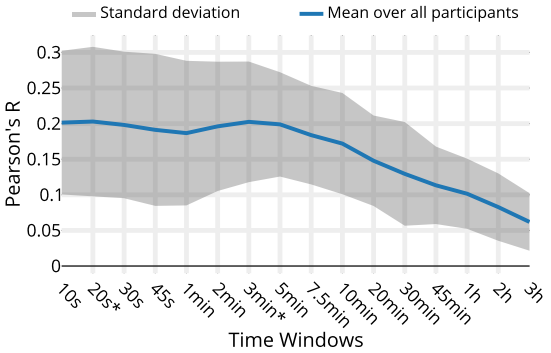
the tasks used in these studies. To determine the optimal time window(s) for predicting interruptibility, we used the following three commonly used statistics to calculate a set of four metrics for each combination of time window and feature: (1) Pearson's R measuring the linear correlation between two variables, (2) ANOVA's F measuring the ratio of the between-group variability to the within-group variability, and (3) Mutual Information (MI), a measure that is linked to the concept of entropy and captures the amount of information obtained about one random variable through observing the other random variable. We chose these three statistics to capture a broad range of possible dependencies between the outcome measure and the feature values with Pearson's R and ANOVA's F capturing linear relationships and MI to capture non-linear dependencies. Based on the three statistics, we calculated a total of four metrics since we calculated the F-score for both classifications (two states of interruptibility) using `f_classif` and regression (7 states of interruptibility) using `f_regression`. As we did not have enough samples to compute MI for classification reliably, we only computed it for regression using `mutual_info_regression`. We used scipy.stats [Jones et al., 01] to calculate Pearson's R and scikit-learn [Pedregosa et al., 2011] for the other metrics.

We visually inspected the graphs that we generated for each feature, metric and each participant (see an example in Figure 3.3 (a)) and found that the line graphs from different participants have similar trends and slopes (see Figure 3.3 (a)). We therefore aggregated the data from all participants by calculating the mean and standard deviations of each metric's absolute values and generated a graph for each feature (see an example in Figure 3.3 (b)). Finally, we compared the four different metrics with each other by generating graphs for each feature including all metrics (see an example in Figure 3.3 (c)). We found that all four metrics were highly correlated, even the mutual information metric (Pearson's R and `f_classif`: Pearson r=.92, p<.000001, Pearson's R and `f_regression`: Pearson r=.95, p=.0, Pearson's R and `mutual_info_regression`: Pearson r=.84, p<.000001) and that they have similar peaks (see Figure 3.3 (c)). We ended up choosing the time windows that maximized the absolute mean of Pearson's R over all participants through manual visual peak detection. When there were

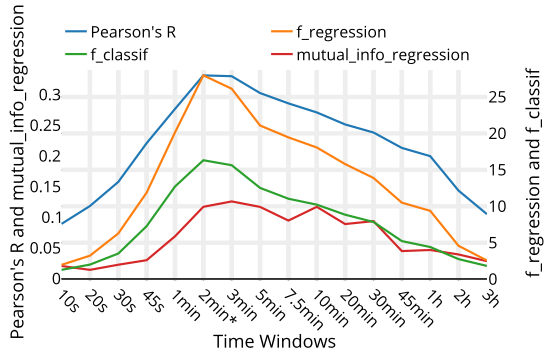
several peaks or in the rare cases where the metrics had substantially different peaks (e.g. due to a non-linear dependency), we added each peak as a time window. The latter occurred for 15 of the 55 features for which we determined a time window.



(a) Pearson correlation between interruptibility ratings and the feature *percentage of time spent in software development* over all time windows and per participant.



(b) Overall Pearson correlation between interruptibility ratings and the feature *Polar mean of HR* extracted over all time windows.



(c) All four metrics used to compare time windows for predicting interruptibility using the feature *number of steps per min* averaged over all participants.

Figure 3.3: Selection of graphs generated to determine the optimal time window for predicting interruptibility (chosen time window denoted with *).

The selected time windows per feature are listed in (blue) in Table 3.2. For the biometric features (heart and movement), shorter time windows between 10s and 3min were generally better than longer ones, whereas for user input and application window features, longer windows between 10min and 20min were better. Exceptions were communication (3h) and software development (2min). Our results show that the optimal time window varies per feature and suggest a range of time windows which work well for certain feature groups.

3.5.2 Sensors, Features and Perceptions

To evaluate the accuracy of predicting interruptibility in the field and compare the predictive power of the various features, we applied machine learning to the collected features as well as groups of features. To add to the understanding of participants' perception on interruptibility and in particular how and why specific features might relate to their interruptibility, we further complement the quantitative findings with an analysis of our diary survey and interviews.

Interruptibility Prediction

The goal of our research is to predict a person's interruptibility in a specific moment with high accuracy using the features extracted from the collected biometric and computer interaction data. We use the ratings from the participants' experience samples split into two states as ground truth. Since biometric and computer interaction data is highly individual and trained models can often not easily be transferred to new participants [Fritz et al., 2014, Züger and Fritz, 2015, Visuri et al., 2017], we trained models individually for each participant, similarly to Haapalainen et al. [Haapalainen et al., 2010]. For each participant, we predicted interruptibility using ten trials of stratified ten-fold cross-validation, which keeps the class proportions consistent in each fold, and a random forest classifier pipeline (500 estimators) with initial feature imputation and standard scaling.

Table 3.3 presents the accuracy scores for each sensor and combinations thereof. As baseline accuracy we report the accuracy that a majority classifier

	Interruptibility Prediction Accuracy (2 States, Individual Models)													
	P01	**P02**	**P03**	**P04**	**P05**	**P06**	**P07**	**P08**	**P09**	**P10**	**P11**	**P12**	**P13**	**All**
Baseline Accuracy	66%	63%	53%	58%	58%	53%	71%	82%	53%	61%	52%	56%	57%	**60.2%**
Fitbit	64%	72%	72%	64%	66%	57%	65%	79%	61%	74%	63%	65%	59%	**66.2%**
Polar	66%	67%	56%	59%	58%	55%	69%	77%	73%	62%	59%	54%	59%	**62.5%**
Computer Monitoring	78%	69%	80%	73%	70%	74%	74%	85%	85%	76%	74%	72%	62%	**74.8%**
Fitbit + Polar	68%	76%	70%	65%	61%	58%	69%	81%	72%	76%	67%	63%	61%	**68.3%**
Fitbit + Computer Monitoring	79%	73%	80%	75%	70%	74%	74%	86%	85%	78%	74%	72%	64%	**75.7%**
Polar + Computer Monitoring	78%	72%	80%	74%	69%	72%	73%	85%	86%	77%	73%	73%	62%	**75.0%**
Fitbit + Polar + Computer Monitoring	79%	76%	79%	74%	69%	72%	74%	85%	86%	78%	74%	72%	62%	**75.3%**

Table 3.3: Prediction results using different sensors and combinations thereof per participant and averaged over all (the darker the color the higher the accuracy).

would achieve that always predicts the class containing more samples. The results are obtained training individual models for two states of interruptibility. While all sensors were better than the baseline, the features of the computer interaction sensors (accuracy=74.8%) were more predictive compared to the features from the biometric sensors (accuracy=68.3%). Adding one or both biometric sensors slightly improves the classifier (accuracy=75.7%). When comparing the Polar and the Fitbit sensors, for 9 of 13 participants the Fitbit yielded better results, while for the remaining 4 the Polar was more accurate (accuracy Fitbit = 66.2%, accuracy Polar = 62.5%). Note that the Fitbit comprises a wider variety of features, e.g. step count, than the Polar.

Table 3.2 contains the feature importance attributed by the random forest classifier using all features and averaged over all participants' individual models. For the feature importance metric we used the Gini impurity measure from scikit-learn [Pedregosa et al., 2011] that is attributed to each feature by the random forest classifier and captures the feature's ability to avoid misclassification [Rokach and Maimon, 2005]. The most important features are the application window group and user input, followed by heart and sleep measurements. Calendar (2.8%), movement (2.3%) and time related features (2.1%) are the least important, contributing only 7.2%.

Developers' Perceptions of Interruptibility

To complement our quantitative comparison of features and sensors, we analyzed the interview and diary survey data to learn more about how software developers perceive interruptibility and related factors, and whether their perception matches our feature model. We analyzed the interview audio recordings by transcribing and applying open and axial coding and the diary survey data using multiple regression analysis.

Similar to our classification results that show that application window and user input features are most predictive, all participants stated in the interview that their interruptibility changes with certain activities on the computer, such as coding or writing emails, but only a few also explicitly mentioned the user input (15% of participants).

" When I do development or code reviews I am very focused and not interruptible. During email writing on the other hand, I am more interruptible." (P04)

"If I am typing something, sure I might forget what I was typing when I get interrupted." (P12)

In addition and consistent with prior work (e.g. [Bailey and Iqbal, 2008]), participants stated that they are more interruptible at task boundaries (69%). While this is not explicitly covered in our examined features, this is somewhat implicitly captured by user input and application window features while participants are working at the computer as previous research has shown [Tanaka and Fujita, 2011, Nair et al., 2005, Iqbal and Bailey, 2008], as well as with the features related to being idle, calendar entries and physical movement, e.g. when changing location and coming back from a meeting [Ho and Intille, 2005, Fisher and Simmons, 2011, Komuro et al., 2017, Stern et al., 2011].

"... [I am more interruptible] between tasks, when I organize myself and plan my next step." (P03)

"... [more interruptible] around meetings, because it takes me a bit of time to get back in the flow of things." (P09)

Participants further mentioned that their interruptibility depends on internal states such as sleepiness (85%), focus (77%), mood (46%), challenge (38%),

productivity (38%), stress (38%), health (23%), and engagement (15%).

"When [last night] was relatively short, I have a hard time to concentrate anyways, and want to be disturbed less." (P03)

"When I am kind of frustrated or nervous, I am more annoyed if someone interrupts me." (P13)

"When I was doing a complicated code review, where I first had to understand the dependencies, it would not be good to be interrupted." (P05)

This overlaps with the sensors we chose, especially the biometric ones, as they have the potential to measure a variety of internal states such as stress, mood or mental load [Healey and Picard, 2005, Haag et al., 2004, Müller and Fritz, 2015, Haapalainen et al., 2010]. As an example, we chose the Polar H7 to measure HRV, which is a well-established indicator for stress (e.g. [Melillo et al., 2011]).

When asked about temporal patterns of interruptibility over the course of the day, many participants stated that they do not necessarily think that there is a direct link to interruptibility, but rather that the routine of activities and external factors such as background noise and interruption frequency is linked and might vary throughout the day.

"There is nothing specific about the time of day, it is just how my routine is laid out." (P06)

"Around lunch time is the busiest time of the day." (P02)

Most participants find it easier to focus, which would result in lower interruptibility, when the office is quieter (46%).

"After 5pm many go home and then it's very quiet, then it is easier to concentrate." (P02)

To further examine temporal patterns of interruptibility, we visually analyzed the interruptibility ratings in relation with the time of day. Similar to the interview responses, we could not find any consistent and significant patterns across participants, which is also supported by the fact that time related features were only weighted by 2.1% in the interruptibility classifier.

In our daily diary survey that we performed throughout the study period, we asked participants to rate their relative overall interruptibility for the whole day. We further asked them to rate several features (listed in Table 3.4) that were referenced in prior work in relation to interruptibility and work focus [Rosekind

Obs.: 151, Adj. R^2=.28, R^2=.32, F(8, 142)=8.34, p<.001			
int. frequency*	(β=.14, p=.025)	engagement*	(β=-.41, p<.001)
productivity	(β=-.12, p=.23)	challenge	(β=-.08, p=.44)
stress	(β=-.16, p=.07)	sleepiness	(β=.11, p=.12)
valence	(β=.11, p=.21)	arousal	(β=-.06, p=.45)

Table 3.4: Linear regression results with daily interruptibility as dependent and feature ratings collected in the daily survey as independent variables (* denotes significance at p<.05).

et al., 2010, Bailey and Iqbal, 2008, Müller and Fritz, 2015, Mark et al., 2008, Mark et al., 2014] and that are to some extent captured by our sensors, especially the biometric ones. We found that the interruptibility ratings per day collected with the experience sampling prompts and the daily interruptibility rating from the diary survey correlate significantly (Pearson r=0.42, p<.000001), which provides support for the validity of the measures. We then performed a multiple linear regression analysis with the daily interruptibility rating as the dependent variable using all 151 recorded responses from all participants. The results (shown in Table 3.4) show that participants were more interruptible when they had many interruptions, and less when they were engaged, and that there is a trend (not significant though) that participants were more interruptible when they were sleepy, and less when they were stressed or productive.

Overall, our results indicate that there is a strong overlap between the features determined as particularly predictive in our analysis of the sensor data and the perceptions of participants.

3.5.3 Interruptibility Prediction in the Field

To investigate the general use and sensitivity of our interruptibility classification in the field, we first create and compare a general model trained across several participants with our individually trained models, and second, examine the classification of a more fine-grained interruptibility.

The main advantage of a general model is that no initial training phase is needed to use it on a new subject in practice. For our analysis, we used leave-

	Valid Samples	Skipped Samples	Histogram	Individual Models 2 States Base	2 States Acc.	2 States Impr.	3 States Base	3 States Acc.	3 States Impr.	7 States Base	7 States Acc.	7 States Impr.	General Models 2 States Acc.	General Models 2 States Impr.
P01	217	3		66%	79%	19%	46%	64%	39%	31%	41%	31%	67%	1%
P02	142	2		63%	75%	18%	59%	69%	16%	33%	40%	21%	66%	4%
P03	195	4		53%	80%	50%	53%	67%	27%	23%	39%	71%	82%	54%
P04	200	8		58%	75%	30%	52%	60%	17%	24%	36%	55%	74%	28%
P05	127	1		58%	69%	18%	43%	48%	11%	20%	27%	30%	71%	22%
P06	172	16		53%	73%	36%	40%	60%	51%	28%	38%	38%	64%	20%
P07	135	0		71%	73%	3%	49%	70%	43%	44%	51%	17%	70%	-1%
P08	152	0		82%	85%	4%	65%	73%	12%	33%	48%	46%	67%	-18%
P09	191	0		53%	86%	62%	43%	75%	75%	39%	68%	73%	76%	45%
P10	484	4		61%	78%	28%	56%	77%	38%	51%	69%	36%	64%	5%
P11	162	0		52%	73%	40%	62%	68%	9%	26%	39%	51%	73%	39%
P12	145	29		56%	71%	28%	63%	62%	-2%	29%	32%	10%	73%	31%
P13	193	17		57%	63%	10%	60%	59%	-2%	25%	25%	0%	60%	5%
Totals:	**2515**	**84**		**60.3%**	**75.3%**	**26.6%**	**53.1%**	**65.5%**	**25.7%**	**31.2%**	**42.5%**	**36.9%**	**69.8%**	**18.0%**

Table 3.5: Results for predicting 2, 3 and 7 states of interruptibility along with the size and histogram of the available samples' interruptibility labels. The last column reports results from general models trained on all but one and tested on the one participant. *Legend: "Base": Baseline accuracy obtained by a majority classifier, "Acc.": Accuracy, "Impr.": Percentage improvement over majority classifier*

one-out cross-validation for which we iterated over all participants and trained a classifier with data from all participants except one and tested it on the remaining one [Visuri et al., 2017]. The results show that the general model achieves equal or better accuracy than the baseline for all except the two participants P07 and P08 (see last column of Table 3.5). At the same time, and not surprising, the individual models performed better for almost all participants, except for three (P03, P05, and P12), with a 75.3% averaged accuracy over all participants compared to 69.8% for the general model.

To investigate how many training samples per individual are approximately needed to build an individual interruptibility classifier that is as good or better than the general model, we produced learning curves for each individual using shuffle split cross-validation (100 splits, test size of 20% of the available samples). Figure 3.4 depicts an example of a learning curve for one participant (P06). The illustrated example shows that already with few samples (approximately 40 in this case), the individual classifier starts outperforming the general model and improves with increasing sample size. Over all participants, an average sample size of 20 to 80 was sufficient to train an individual interruptibility classifier that is close or outperforms the general one.

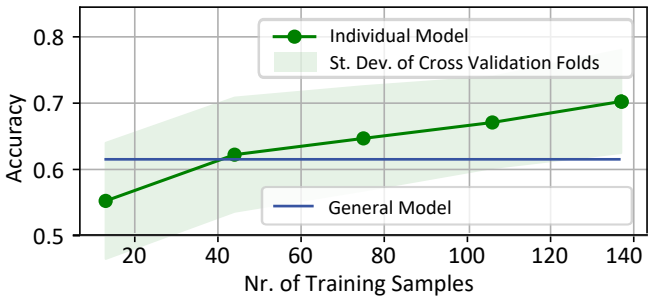


Figure 3.4: Learning curve for participant P06.

In a real-world application, a more fine-grained classification of interruptibility might also be valuable, e.g. when indicated to co-workers, it might enable a more informed decision whether to interrupt someone or not, also with respect to the priority and kind of the interruption. We therefore examined the accuracy of a more fine-grained classification by splitting the outcome measure—the interruptibility rating—into three and seven states. For this analysis, we used the best feature set, i.e. all features, that we determined earlier. Table 3.5 presents the results for interruptibility predictions into several granularities for each participant. Average prediction accuracies were 75.3%, 65.5% and 42.5% for prediction into two, three and seven states of interruptibility, which is an average improvement of 26.6%, 25.7% and 36.9% compared to a majority classifier. The aggregated confusion matrices for prediction into three and seven states reveal that mis-classifications rarely fall into distant classes, but often into adjacent ones (see Table 3.6 for seven states). These results indicate that a classifier trained on the collected computer interaction and biometric features is able to predict interruptibility with reasonable accuracy not only into two, but also three and even seven states of interruptibility in the field.

True Label \ Predicted Label	1	2	3	4	5	6	7
1	15%	2%	1%	1%	0%	0%	1%
2	3%	5%	4%	1%	1%	0%	1%
3	1%	3%	6%	2%	1%	1%	1%
4	3%	2%	4%	5%	1%	1%	1%
5	1%	1%	2%	2%	2%	2%	1%
6	1%	0%	1%	1%	1%	3%	2%
7	1%	0%	1%	1%	0%	2%	10%

Table 3.6: Aggregated (summed up) confusion matrix for seven states from individual models of all participants.

3.6 Discussion

In the following, we discuss our findings, in particular implications from the time window analysis and feature comparison, practical applications of the interruptibility classifier as well as limitations and threats to validity.

Time Windows and Features. Our results suggest that a developer's interruptibility is not only affected by the few seconds and minutes before an interruption, but that there are features, such as the activities or sleep, that can have a longer lasting effect on interruptibility. While most prior work focuses on features calculated for short time windows of up to 5min [Hudson et al., 2003, Vorburger et al., 2011, Züger and Fritz, 2015], we analyzed a wider range of features and time windows spanning from 10s to 3h for most features and a whole day for some, such as sleep and resting HR. Our results show that for certain feature groups, longer time windows are more informative and that even daily features have an importance for predicting interruptibility, e.g. 4.4% importance for sleep (see Table 3.2). For example, communication related activities were most correlated to interruptibility using large time windows of 1 to 3h. This longer lasting effect of communication related activities was also mentioned in a previous study that found that office workers feel less productive after spending a longer amount of time with email activity [Mark et al., 2016b], which in turn might impact their interruptibility. We also found that there were several 'good' time windows

for certain features. A possible explanation is that these time windows refer to different notions of the feature. An example is the *max. time in an activity category*. For a short time window (10s) it might indicate whether the person is at a breakpoint or task switch, while for longer time windows (20min) it is more indicative of extended focus. In general, our findings show that there are certain ranges of time windows for certain feature categories, but that there is not necessarily just one best time window for each feature. Future studies should therefore further analyze how the feature under investigation varies over time.

Sensor Comparison. Previous research has already linked both computer interaction and biometric sensors to mental load and interruptibility [Haapalainen et al., 2010, Komuro et al., 2017, Vidaček et al., 1986, Iqbal and Bailey, 2008, Nair et al., 2005, Fogarty et al., 2005b]. To the best of our knowledge, our study is the first to compare these types of sensors in the field over a longer period of time. While our study demonstrates that computer interaction features can be used to accurately predict interruptibility at the computer and that they are more predictive than the biometric features used in our study, the results also show that biometric sensors already have a great potential in accurately predicting interruptibility in the field despite the noise. For our study, we focused on two biometric sensors that we selected due to their little invasiveness, cost and availability. Especially with the rapid advances in technology in combination with biometric sensors being less limited to a specific workstation and being able to capture more of a person's work day, our results demonstrate the potential of these types of sensors for the future. Overall, participants perceived the computer interaction sensors as less invasive, but thought that the captured data was more sensitive than the biometric data in the work context. Biometric sensors can thus serve as a complement or substitute to improve accuracy, or respect privacy preferences for now.

Practical Use. Our findings show that using a general interruptibility classifier is accurate enough to successfully break the cold start problem. For practical use, we therefore suggest using a general model as a default and allowing the

user to improve the classifier by training it. Even with few individual samples one is able to achieve a high accuracy with this approach. In general, such a classifier can then be used to indicate a knowledge worker's interruptibility to potential co-workers, which has been explored with physical indicators [Züger et al., 2017], or indicators displayed on the computer [Begole et al., 2004, Lai et al., 2003, Fogarty et al., 2004]). Similarly, such an interruptibility classifier can be used to mediate interruptions directly by postponing computer-based interruptions while a person is non-interruptible to a more opportune moment, which has also been investigated in prior work [Iqbal and Bailey, 2008, Horvitz et al., 2004, Kapoor and Horvitz, 2007, Kapoor and Horvitz, 2008, Horvitz et al., 2002]. A further potential use of the data is to display the current and historical interruptibility state to the knowledge worker herself. Given the strong links between interruptibility and states such as focus or stress, increased awareness about one's interruptibility patterns might help knowledge workers to reflect on their work patterns and potentially improve their work experience. Several of our participants already enjoyed the biometric data by itself a lot.

Limitations. We conducted our study with software developers working in offices, which limits our results to this context. While we can assume that the results can be generalized to similar job roles and environments, more research needs to be conducted to study interruptibility in other areas. We further prompted our participants to rate their interruptibility using a pop-up displayed on the computer, which limits the times of responses to times spent at the computer. Therefore, we were not able to collect self-reports during times spent away from the computer. However, some of our features (e.g. heart and movement data) were collected at all times, even when the participant was away from the computer, and we have several data points from prompts that were answered shortly after returning to the computer. In fact in 17% of our data samples participants answered the prompt less than 1 minute after an idle period without computer interaction. Also, the high predictive power of the feature group "Activity Category" might be partially due to the manual labeling of applications into categories (e.g. *Visual Studio* into *Software Development*) which includes

fine-grained expertise of the annotators. However, once the mapping exists, the categorization can be completely automated. When the the set of used application changes, a manual update of the mapping would be necessary to ensure a sustained high predictive power of this feature group.

Threats to Validity. The interruptibility rating pop-up is, ironically, an interruption in itself and could have potentially disrupted our participants in their work flow. As participants usually only needed a very short time to answer the prompts (in 53% of all cases the pop-up was answered within 10s) and as they only rarely postponed a prompt (3% of all prompts), we are confident that the pop-ups did not disrupt our participants from their work flow notably. Another threat to validity is that participants might not be able to assess their interruptibility correctly or that they might not have understood the question. We ensured that we spent enough time to explain the pop-ups at the beginning of the study to mitigate this risk. Furthermore, not every one of the collected samples in our dataset contains full data from all sensors, which might influence the comparison of the sensors and features, e.g., computer interaction data is inherently limited to times spent at the computer. Missing values were imputed by replacing with the mean before classification, as this technique can lead to better results than discarding them which would decrease the sample size.

3.7 Conclusion and Future Work

In this chapter, we presented the results of a two-week field study with 13 professional software developers in which we examined the use of a wide variety of biometric and computer interaction features to predict interruptibility. Our analysis shows that we are able to predict interruptibility at the computer with 75.3% accuracy (a 26.6% improvement over the baseline) and that computer interaction features are more accurate than the biometric ones (74.8% vs. 68.3%). We further show that the best time windows to extract features vary across feature categories and that certain features can affect interruptibility over long periods of time. Finally, we show that even a generally trained model can

accurately predict interruptibility for new subjects to overcome the cold start problem, and that even small sets of samples can be used to rapidly improve the classifier.

As a next step, we plan to generalize our model to a broader range of knowledge workers and explore its potential to actively reduce interruption cost by indicating the interruptibility status to co-workers and fostering undisrupted work.

3.8 Acknowledgments

The authors would like to thank all study participants. This work was funded by SNF.

4

Reducing Interruptions at Work: A Large-Scale Field Study of FlowLight

Manuela Züger, Christopher Corley, André N. Meyer, Boyang Li, Thomas Fritz, David Shepherd, Vinay Augustine, Patrick Francis, Nicholas Kraft, Will Snipes
Published at the 2017 CHI Conference on Human Factors in Computing Systems

Contribution: minor part of the tool development (focused on the automatic status update algorithm), major part of the study design, minor part of the data collection (installed the FlowLight on 4 sites in 3 countries), major part of the data analysis and paper writing

Abstract

Due to the high number and cost of interruptions at work, several approaches have been suggested to reduce this cost for knowledge workers. These approaches pre-

dominantly focus either on a manual and physical indicator, such as headphones or a closed office door, or on the automatic measure of a worker's interruptibility in combination with a computer-based indicator. Little is known about the combination of a physical indicator with an automatic interruptibility measure and its long-term impact in the workplace. In our research, we developed the FlowLight, that combines a physical traffic-light like LED with an automatic interruptibility measure based on computer interaction data. In a large-scale and long-term field study with 449 participants from 12 countries, we found, amongst other results, that the FlowLight reduced the interruptions of participants by 46% (based on 36 interruption logs), increased their awareness on the potential disruptiveness of interruptions (based on 183 survey responses and 23 interview transcripts) and most participants never stopped using it (86% of the 449 users continued even after the end of the two-month study period).

4.1 Introduction

Knowledge workers are frequently interrupted by their co-workers [González and Mark, 2004, Czerwinski et al., 2004, Sykes, 2011]. While many of these interruptions can be beneficial, for instance to resolve problems quickly [Isaacs et al., 1997], they can also incur a high cost on knowledge workers, especially if they happen at inopportune moments and cannot be postponed [Bailey and Konstan, 2006, Mark et al., 2008, Borst et al., 2015, McFarlane, 2002].

Due to the high cost and the high number of interruptions that knowledge workers experience every day (e.g., [Czerwinski et al., 2004, González and Mark, 2004]), several approaches have been proposed that can roughly be categorized by the interruptions they address: computer-based and in-person. Studies have shown that the cost of computer-based interruptions can successfully be mitigated by automatically detecting a knowledge worker's interruptibility and mediating interruptions by deferring them to more opportune moments (aka. defer-to-breakpoint strategy) [Iqbal and Bailey, 2008, Arroyo and Selker, 2011, Ho and Intille, 2005]. Another strategy to reduce the cost of computer-based interruptions is to indicate a person's interruptibility to co-workers in a contact-list style

application on the computer [Tang et al., 2001, Lai et al., 2003, Begole et al., 2004]. While these approaches have also been suggested for addressing in-person interruptions, they did not show to have any effect on them, probably since the contact-list style applications can easily be hidden behind other applications and thus forgotten at communication initiation [Lai et al., 2003, Begole et al., 2004, Fogarty et al., 2004, Hincapié-Ramos et al., 2011a].

For in-person interruptions—one of the most costly kind of interruptions due to their high frequency and immediate nature [Sykes, 2011, González and Mark, 2004, McFarlane, 2002]—approaches predominantly rely on manual strategies to physically indicate interruptibility, such as wearing headphones, closing the office door, or using busy lights that have to be set manually [Sykes, 2011, Embrava, 2016]. Since manual approaches are cumbersome to maintain, users generally don't update them on a regular basis and their accuracy and benefits are limited [Milewski and Smith, 2000]. Only very few approaches have looked at a combination of a physical interruptibility indicator with an automatic interruptibility measure to reduce the cost of in-person interruptions [Hincapié-Ramos et al., 2011b, Bjelica et al., 2011] and there is no knowledge on the long-term effects of such approaches.

In our research, we developed the FlowLight approach, an approach to reduce the cost of in-person interruptions by combining a physical interruptibility indicator in the form of a traffic-light like LED (light emitting diode) with an automatic interruptibility measurement based on a user's computer interaction. In a large-scale and long-term field study with 449 knowledge workers from 12 countries and 15 sites of a multinational corporation, we evaluated the FlowLight and its effects in the workplace. Over the course of the study, we collected a rich set of quantitative and qualitative data, including self-reported interruption logs of 36 participants, survey responses of 183 participants that used the FlowLight for at least 4 weeks, and in-depth interviews of 23 participants. Our analysis of the data shows, amongst other results, that the FlowLight significantly reduced the number of interruptions of participants by 46%, while having little impact on important interruptions. Further, the FlowLight increased the awareness on the cost of interruptions within the workplace, participants felt more productive using

the FlowLight and 86% of the 449 participants continued using the light even after the two-month study period ended. Overall, the gained insights on the long-term usage of the FlowLight provide strong support for the benefits of combining a physical interruptibility indicator with an automatic interruptibility measure in the workplace and its significant impact on reducing in-person interruption costs.

4.2 Related Work

Related work on managing interruptions can broadly be grouped into strategies for reducing interruptions and disruptiveness, and ways of measuring and indicating interruptibility.

4.2.1 Reducing Interruptions and their Disruptiveness

Knowledge workers have long recognized the detrimental effects of interruptions and have sometimes developed their own techniques for managing them. These techniques include the use of instant messaging to negotiate availability for an interruption beforehand and reduce the disruptiveness for the interrupted person [Nardi et al., 2000], as well as the use of manual and physical indicators, such as headphones or a closed office door to either signal unavailability or tune out distractions [Sykes, 2011].

In addition to these informal means, researchers have developed approaches to reduce the negative effects of interruptions. One strategy to reduce the disruptiveness of interruptions is to defer them from moments when the interruptee is in the middle of a task to naturally occurring breakpoints—aka. 'defer-to-breakpoint' strategy. This idea is based on studies finding that the cognitive load drops at task boundaries, and that interruptions at lower cognitive load are less harmful [Bailey and Iqbal, 2008, Borst et al., 2015]. Iqbal and Bailey developed a system that implements a defer-to-breakpoint policy to reschedule notifications to more opportune moments and found that they caused less frustration and shorter reaction times [Iqbal and Bailey, 2008]. Ho and Intille used accelerometers

to detect activity transitions and found that messages on mobile devices were better received during transitions compared to random times [Ho and Intille, 2005]. While these approaches have been successful at mitigating interruptions from the computer and mobile devices, they do not address the frequent and costly in-person interruptions in workplaces that the FlowLight targets.

A second strategy that builds upon the idea of deferring interruptions to more opportune moments is to indicate a knowledge worker's interruptibility to potential interrupters and thereby implicitly help negotiate the timing of the interruption. In the following, we discuss approaches to measure and to indicate a knowledge worker's interruptibility.

4.2.2 Measuring Interruptibility

Previous research has explored various features to measure a person's interruptibility. For instance, Hudson et al. simulated sensors by coding audio and video recordings into features related to the person's current context, such as the number of people present or the phone being on the hook [Hudson et al., 2003]. While their approach showed promise in measuring interruptibility, the chosen features are difficult to capture automatically.

To automatically detect a person's interruptibility, Stern et al. developed an approach that is based on the person's location and calendar information [Stern et al., 2011]. Fogarty et al. used speech sensors, location and calendar information and activity on the computer to measure presence and availability [Fogarty et al., 2005b]; Tani and Yamada measured interruptibility using the pressure applied on the keyboard and mouse [Tani and Yamada, 2013]; and Coordinate by Horvitz et al. uses user activity and proximity of multiple devices to forecast presence and availability [Horvitz et al., 2002].

More recently researchers have also started to use biometric data to measure interruptibility. For instance, Kramer classified interruptibility during a US military training with an electroencephalography (EEG) sensor that captures the electrical activity of the brain [Mathan et al., 2007]. Chen et al. calculated interruptibility based on an electromyography (EMG) sensor that captures heart rate variability and muscle activity [Chen et al., 2007]. In our previous work,

we used various biometric sensors (EEG, electrodermal activity (EDA), skin temperature, and photoplethysmography (PPG)) to predict interruptibility [Züger and Fritz, 2015]. Overall, research has shown that biometric sensors can be valuable in automatically measuring interruptibility, however, at this point the biometric sensors required to accurately measure interruptibility are generally still too invasive for long-term usage.

The FlowLight builds upon previous research in this area by automatically measuring interruptibility based on a combination of computer activity, calendar information and log-in state. It thereby utilizes a minimally invasive set of features that performs well without compromising the users' privacy or requiring additional body-worn biometric sensors. It further extends previous research in this area by combining the automatic measure with a physical indicator.

4.2.3 Indicating Interruptibility

To indicate a knowledge worker's interruptibility to co-workers, most prior research focused on contact list-style tools that are installed on the user's computer and vary mostly in the data that is used to determine availability/interruptibility. For instance, the ConNexus tool has a contact list view that provides awareness information on a person's device idleness, log-in state and activity history and thus indicates a person's availability to facility communication for the integrated communication channels, such as IM [Tang et al., 2001]. Awarenex and Lilsys build on ConNexus, adding mobile location tracking and physical presence sensors, respectively. An evaluation of these tools found a qualitative improvement in interruption awareness but no reduction in the number of interruptions [Begole et al., 2004]. Lai et al.'s MyTeam approach uses information on presence, network connection and mouse and keyboard activity to indicate availability in a contact list. In a small user study, they found that the approach decreased the number of phone calls and voice mails but increased the face-to-face interruptions [Lai et al., 2003]. Fogarty et al. developed MyVine that integrates with a phone, IM and an email client and uses context information from speech sensors, computer activity, location and calendar information. A four week study revealed that the context information was mainly used as presence indicator and did not prevent

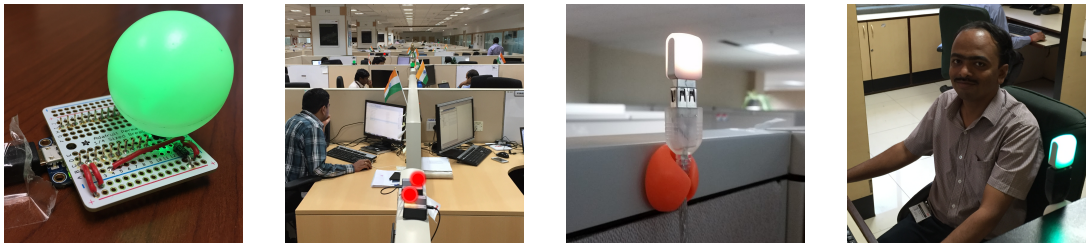
interruptions via IM [Fogarty et al., 2004]. Overall, study results for these computer-based interruptibility indicators suggest that they can help increase awareness on the disruptiveness of interruptions, which could be a first good step as stated by Beyea [Beyea, 2007]. However, the results also suggest that these approaches do not reduce in-person interruption costs, which is what the FlowLight is addressing.

Since in-person interruptions are one of the top causes for interruptions in the workplace and their immediate nature makes them particularly disruptive [Sykes, 2011], researchers found that knowledge workers use physical indicators, such as headphones or office doors to indicate interruptibility and reduce interruptions and distraction [Sykes, 2011].

Only few researchers examined indicators that are not just visible on a knowledge worker's computer monitor. InterruptMe projects availability cues of possible contacts onto a wall at the time when the interrupter is about to initiate a communication [Hincapié-Ramos et al., 2011b, Hincapié-Ramos et al., 2011a]. The MoodLight uses an ambient display connected to an electrodermal activity (EDA) monitor that indicates the excitement level of one or two individuals [Snyder et al., 2015]. Bjelica et al. developed an automatic interruptibility indicator that displays the status through ambient lighting effects and found in a small and short study that the indicator reduced the number of interruptions [Bjelica et al., 2011]. The FlowLight presented in this paper uses a physical traffic-light like LED placed at the desk of each person, such that the person's interruptibility status can be seen by anyone approaching. Thereby, our approach is more direct and prominent than subtle ambient lighting and different to previous research, our large-scale field study examines the long-term effects of such physical indicators.

4.3 Approach and Implementation

The FlowLight consists of a computer application to automatically determine a user's interruptibility state and a physical LED light to indicate this state to co-workers. The FlowLight was developed iteratively over more than a year and



(a) July 2015: Original prototype

(b) August 2015: First major pilot (lights emphasized with overlays)

(c) October 2015: Blink(1) version with adhesive clip

(d) April 2016: second pilot in India (India #2)

Figure 4.1: Evolution of the physical indicator of the FlowLight over time

improved continuously based on feedback from a small developer team that we used for testing, and later on, also based on feedback from study participants.

Physical LED Light. FlowLight uses a physical traffic-light like LED to indicate the interruptibility status to co-workers. This light has evolved throughout the pilots[1]. The first model, which was designed and soldered in-house, is shown in Figure 4.1a. In Figure 4.1b the same model light is shown encased in plastic and deployed in an open office space. Finally, Figure 4.1c shows the blink(1)[2] LED light that we adopted to avoid installation issues with certain drivers immediately after the first major pilot, which was also the first of two pilots in India (denoted as India #1 in Figure 4.2). Typically, we mounted the LED light on a user's cubical wall or outside a user's office.

The light uses different colors to indicate four states: *Available* as green, *Busy* as red, *Do Not Disturb (DnD)* as pulsating red, and *Away* as yellow. Note that these states and colors mimic the ones used by prominent instant messaging services, in particular the one used by the company under study.

Application. The application features three main components: a *Tracker* to capture events relevant for calculating the interruptibility state, a *Status Analyzer*

[1]We use the term *pilot* to refer to each individual field study trial with a separate team.
[2]https://blink1.thingm.com/

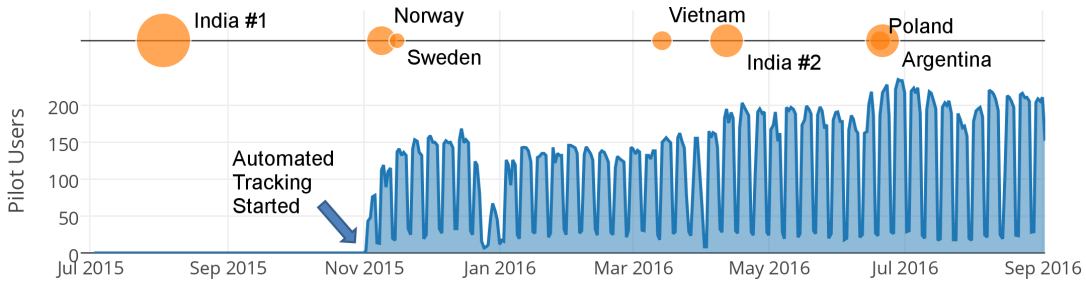


Figure 4.2: FlowLight Users over time (size of orange circles indicates the number of participants; regular dips in the number of users represent weekends and the prolonged dip in December/January 2016 represents the Christmas break)

to analyze the captured events and calculate the user's interruptibility state on the fly, and a *Status Manager* to manage the user's current status, propagating it to the LED light and other applications, in particular instant messaging (IM) clients. The application was implemented to be compatible with the Windows operating system, Skype for Business, an IM and video-conferencing system, and Office 365, a software suite that provides email and calendaring services, amongst others. We chose to tailor our application to these systems and applications due to the IT setup at the target company for our study.

The *Tracker* logs a user's mouse and keyboard interaction. In particular, it collects mouse clicks, movements as pixels moved, scrolling as pixels scrolled and keystrokes (without recording the specific key). This component also logs calendar events to determine meetings and the Skype status.

The *Status Analyzer* uses the tracked keyboard and mouse events to calculate the user's interruptibility status on the fly, i.e., whether the user is available, busy, highly busy (DnD) or away. The algorithms used to calculate the interruptibility status are described below.

The *Status Manager* is notified by the Status Analyzer at every change in the user's interruptibility, and then propagates the updated status to the physical LED light and the user's presence status in Skype for Business. The presence status in Skype for Business can also be changed manually by the user, or automatically by the Office 365 calendar, in case a meeting is scheduled. In

case the presence status is changed manually, the Status Manager updates the interruptibility state of the application and the physical LED light.

Algorithms for Status Updates. Over the course of this study, we used three different algorithms to determine and update the interruptibility status automatically, improving them based on critical user feedback as discussed below.

FlowTracker. This algorithm sums up the computer interaction in the past three minutes according to heuristic weights assigned to each type of event, which were tuned based on feedback from early alpha and beta users of the FlowLight. If the value of the sum is in between the top 9% and the top 4% of their activity range—we captured averages over the past days—the user is considered busy. If it is within the top 4%, the user is considered highly busy. In our first pilot study in Bangalore, India (India #1 in Figure 4.2), we used different thresholds at first, namely 13% and 5% based on a prior study that indicated that knowledge workers are not interruptible for approximately 18% of their day. However, several technical writers (and others) involved in that pilot gave strong feedback that the light switched to the busy state too easily, which is why we lowered the thresholds to the mentioned 9% and 4%.

Smoothing. While the FlowTracker showed promise, many early users complained that it was too sensitive to certain input. For instance, a twenty second burst of typing may cause a user to temporarily be shown as busy. Therefore, the Smoothing algorithm marks users as busy if they were active in each of the last three minutes and exceeded a threshold of 100 combined mouse clicks and key presses in the recent past (between 4 and 7 minutes ago). This algorithm reduces frequent changes by requiring over three minutes of activity to become busy and, once busy, by requiring only one above-threshold minute in the recent past to remain busy. To achieve the highly busy status, users had to be busy at the current point in time and had to be above-threshold for fifteen of the last thirty minutes.

Smoothed FlowTracker. While the Smoothing algorithm leads to fewer status changes, since it relied on a static threshold (i.e., 100 combined mouse clicks and key presses), it did not adapt to individual users' work patterns. For instance,

designers working on drawings tended to use mouse clicks almost exclusively, which makes it difficult to exceed the threshold. Thus, we finally combined the FlowTracker algorithm with the Smoothing algorithm to achieve the advantages of both approaches. This algorithm, currently in use, operates as the Smoothing algorithm, but instead of using a static threshold, it utilizes the FlowTracker algorithm to determine above threshold values. This algorithm eliminated all of the most common complaints reported by pilot users. Further refinement of the algorithm is left for future work.

Although our main intent was to use an algorithm to infer interruptibility, we offered participants a "Manual Only" mode since it was requested by some participants, especially those with management roles that needed to be available to others most of the time, and we noticed (and our study confirmed) that our algorithms might not be accurate for everyone or for all activities requiring focus, such as reading or thinking.

4.4 Evaluation

To evaluate the FlowLight, in particular the combination of the physical indicator and the automatic interruptibility measure as well as its effect on knowledge workers, we conducted a long-term and large-scale field study with 449 knowledge workers. For this study, we installed the FlowLight at over 15 locations in 12 countries of one multinational corporation. Over the course of the study, we collected a rich set of data using a combination of experience sampling, a survey, an interview and computer interaction monitoring. Figure 4.1 illustrates a few pictures of the FlowLight in use in different pilots. Figure 4.2 indicates the increasing and continuous number of participants and the major pilots of this study since its beginning and up to September 2016.

4.4.1 Study Procedure

For each team participating in our field study we conducted the same five-week pilot procedure as illustrated in Figure 4.3. Prior to the start of a pilot, we

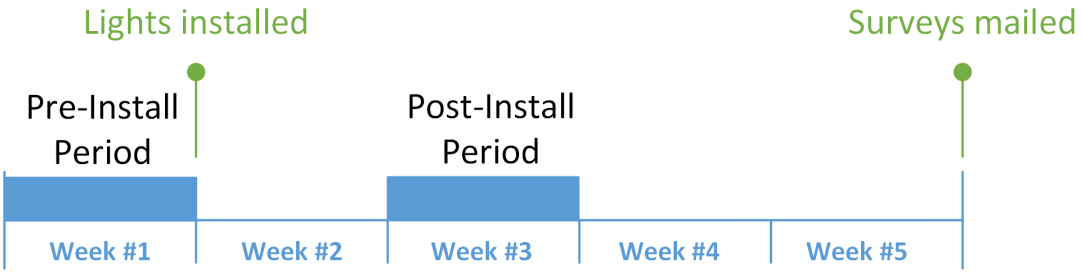


Figure 4.3: Timeline of study procedure

asked the participants to install the FlowLight application in 'data collection only'-mode.

In *Week #1* of the pilot, users were instructed to use the FlowLight application to manually log the time and severity of each interruption during the five day work week. Our application allowed participants to log interruptions by a click on the taskbar menu or a hotkey combination for minimal invasiveness. As soon as an interruption had been logged, a single modal dialog appeared that asked participants to specify the severity of the interruption on a 5-point Likert scale.

At the beginning of *Week #2*, the physical indicator of the FlowLight was installed and the automatic status update feature for the interruptibility status was activated, and we instructed the participants on how to use the FlowLight. To minimize Hawthorne-type effects and have participants and co-workers get used to the FlowLight, we then waited for one week before we gave further instructions [McCarney et al., 2007].

At the beginning of *Week #3*, we again asked participants to manually log their interruptions for 5 work days. We also reminded participants about the manual logging in Week #1 and #3 to ensure they would not forget.

During *Week #4 and #5* users continued using the FlowLight. Throughout these 5 weeks the application collected anonymized usage data. At the end of *Week #5*, after participants had the FlowLight for four weeks, our application prompted them to complete a survey. The survey took an average of 14.2 minutes to complete and had questions on the FlowLight approach and its impact, in particular on participants' interest in continuing using the approach, its impact

on interruption costs, productivity and interaction behavior, on the accuracy of the automatic state detection and manual setting, as well as on general feedback and demographics. After completing the survey, users were asked on the last page of the survey to upload their data collected by the FlowLight application, which included the usage data logs and the logs of the manually captured interruptions.

For a deeper understanding of the long-term usage, experience and effect of the FlowLight, we conducted in-depth interviews with a subset of participants approximately two months after they installed the FlowLight. Interview participants were selected semi-randomly, based on accessibility, availability and willingness to participate in the interview. The interviews were on average 19.5 minutes long and the questions focused on the benefits and limitations participants observed with the FlowLight approach, as well as on how it impacted their own behavior and interactions in the team over the course of the two months since the installation. For instance, we asked participants whether they felt that their colleagues respected their FlowLight or if they noticed situations in which the status was not accurate. Note that the interview and survey questions can be found on our supplemental materials site [3].

Independent of the timeline of the study procedure, we also started to anonymously log the number of people running the FlowLight application each day. For privacy reasons, we only keep track of the number of unique active FlowLight users in the online log.

4.4.2 Participants

Since the beginning of the evaluation, we installed the FlowLight approach with a total of 449 participants from 15 sites, located in 12 different countries, of one multinational corporation. From these 449 participants, we were able to gather:

Survey responses from 183 participants (IDs: S1-S183), 144 male and 39 female, with an average age of 36.0 years (standard deviation, in the following denoted with ±, of 8.7), an average professional experience of 12.0 years (± 8.0), from a variety of work areas, including 77 participants in development, 56 in

[3]https://sites.google.com/site/focuslightonline

other engineering, 24 in project management, 15 in other non-engineering, and 11 in testing, and with various job roles, including 70 individual contributors, 36 other, 32 leads, 31 managers, 8 executives, and 6 architects;

Interview transcripts (conducted by us) from 23 participants (IDs: I1-I23), 22 of which were male, 1 female, average age of 36.9 years (± 5.8), average experience of 13.2 years (± 4.7), and with various job roles, including 9 managers, 11 software developers, 1 researcher, 1 product owner, and 1 tester;

Interruption logs (self-reported) from 36 participants across six different countries, 13 from Argentina, 6 from Norway, 5 from Poland, 5 from Switzerland, 5 from Sweden, and 2 from the USA;

Usage data logs from 47 participants (IDs: D1-D47) 20 from Argentina, 18 from India, 4 from Poland, and 5 from Vietnam.

Online logs from all 449 participants that installed the approach (each one had the application running for at least one day after we integrated the logging feature).

Note that due to privacy concerns with the collected data, we did not require participants to identify themselves in each step and/or fill in their demographics, except for the survey, which is why we can only report some demographics for each round and are not able to track the participants across the different methods, for instance the survey and the self-reported interruption logging.

4.4.3 Data Collection and Analysis

Survey and Interview. In total, we collected survey responses from 183 participants after they had been using the FlowLight for at least four weeks, and interview transcripts from the 23 participants after they had been using the FlowLight for approximately two months. To analyze the textual data of the survey and interview responses, we used techniques based on Grounded Theory, in particular open coding and axial coding to determine higher level themes. To establish a common set of codes and themes, two of the authors applied open axial coding to the same subset of interview transcripts and then established a common understanding and defined a structure for the most commonly men-

tioned concepts. As the topics of the survey and interviews overlap, we used and extended the same coding scheme to analyze the textual survey responses. To validate the analysis of the survey results, two additional authors extracted their main findings from a subset of the responses independently.

Interruption Logs. Interruption logs capture the self-reported interruptions per participant logged with the FlowLight software. We collected interruption logs with at least two logged interruptions from 102 participants. We down-selected these to 36 logs by applying strict filtering criteria to ensure data validity as follows. We excluded all interruptions in all logs that were logged during the first five days after the installation of the FlowLight, as interruptions in the period right after the installation are not representative due to Hawthorne-type effects, such as participants getting used to the FlowLight, and co-workers asking curiosity questions [McCarney et al., 2007]. We then excluded all participants, that logged interruptions for fewer than three days in the pre- or the post-installation period. We chose three days as the threshold for each period to ensure a representative sample of work days for comparison without a too strong bias by individual outlier days. Each of the 36 interruption logs captured a combined average of 9.0 work days ($\pm$ 2.2) for pre- and post-period, and contained an average of 28.9 total logged interruptions ($\pm$ 17.0) per participant for the combined time period. We used these interruption logs to compare the impact of the FlowLight on the number of interruptions rated as disruptive by participants.

Usage Data Logs. We captured usage data logs from a total of 179 participants. These logs consist of computer interaction logs, such as mouse and keyboard events, and FlowLight usage data. Since we wanted to analyze user behavior before and after installing the light, we removed any logs that did not include at least two days before and after installing the light. We also excluded logs older than January 2016, as key usage messages were not yet logged by our software, making the analysis infeasible. We ended up with 47 usage data logs containing a total of 1560 work days. These logs consisted of an average of 7.3

work days (±4.2) prior to light installation and 25.9 work days (±14.0) after light installation per participant.

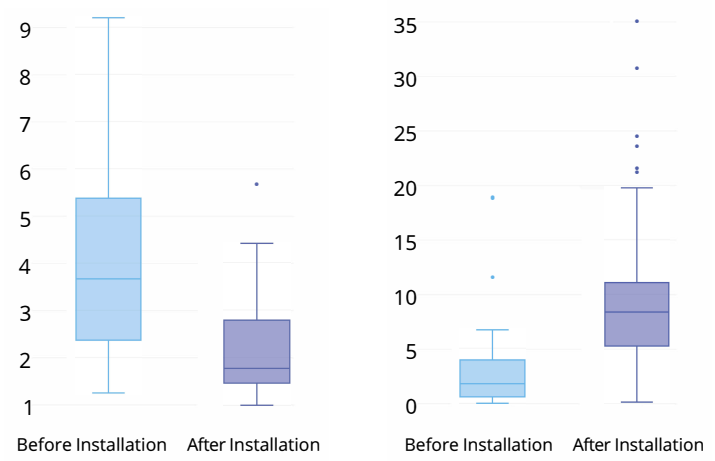
We analyzed usage logs in two ways. First, we counted the number of status change events recorded in the log per day per user for the period before and after the light installation event. It is worth noting that we only included usage logs within the five work days and not on weekends. Second, we used the intervals between status change events detected by one of our algorithms to determine how much time was spent in each status, again for before and after light installation. To eliminate inappropriate intervals (e.g., a user did not turn off the workstation after work), we only accumulated the duration within 12 hours per day.

Online Logs. We collected online logs for a total of 305 days from November 2, 2015 until September 2, 2016 and from 449 participants. These logs were used to determine how many users were using the FlowLight on a given day (as shown in Figure 4.2). We analyzed these logs by summing up the number of unique identifiers that appeared in the log on a given day, which represents the number of active users for that day.

Based on participants' feedback during the period of the field study, we deployed the three main variations of the algorithm described earlier to set the status of the FlowLight. We analyzed differences between the data sets gathered with the three main variations of the algorithms and found no significant differences between the data collected with any two variations, neither in the collected survey items, nor the interruption logs. In the following, we will therefore present the results aggregated over all variations.

4.5 Results

In this section we present the primary findings of our field study. We first examine the effect of the FlowLight on the cost of interruptions before we examine how the FlowLight changed participants' interruption awareness, their interruption-related behavior, and their perception of productivity. Subsequently, we present



(a) Interruptions per participant and day (n=36).

(b) Number of state changes per participant and day (n=47).

Figure 4.4: Logged interruptions and state changes before and after installing the FlowLight.

insights on the costs of the approach, on the influence of its accuracy, on its continued usage by participants and on professional differences.

4.5.1 Reduced Cost of Interruptions

Figure 4.4a is based on the 36 collected interruption logs and illustrates the distribution of the number of interruptions per day and participant in the period before and the period after participants had been using the FlowLight for one week.

Overall, the number of interruptions decreased after the installation and one week usage of the FlowLight by an average of 1.9 (±1.6) interruptions (46%) per participant and day, from 4.1 (±2.1) to 2.2 (±1.1). A Wilcoxon signed-rank test showed that this reduction is statistically significant (Z =-5.0, p <.000001).

A second Wilcoxon signed-rank test only on the number of severe interruptions (disruptiveness rating of 4 or 5) per day and participant further showed that there is also a statistically significant reduction with p <.001 and Z = -3.2.

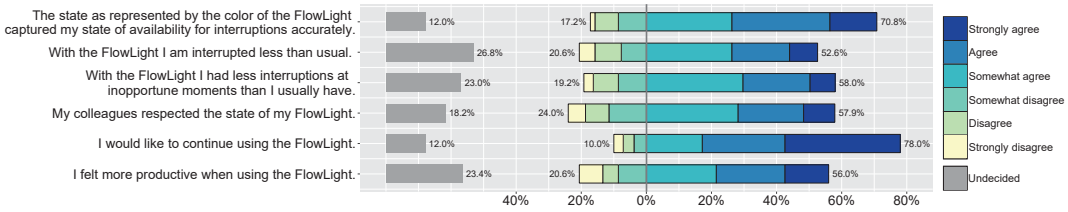


Figure 4.5: Results of a subset of the survey questions (n=183).

An analysis of the survey results (see Figure 4.5 for more detail) further supports that installing the FlowLight reduced the cost of interruptions. 55.0% of the 182 survey participants that answered the question stated that they either strongly agree, agree or agree somewhat that they were interrupted less than usual during their work, while only 20.3% disagreed with it. Even more participants, 59.3%, agreed that they had less interruptions at inopportune moments than usual, whereas only 19.8% disagreed with this statement.

During interviews participants echoed this quantitative evidence. In interview excerpts (full quotes listed in subsequent subsections) participants consistently mentioned that interruptions were reduced. They claimed that the pilot *"..resulted in less interruptions.." (S126)* , eliminated interruptions from colleagues (e.g., *"When [the light]'s red I think they don't interrupt." (I11)*), and *"..didn't stop [interruptions] completely but they surely reduced." (S16)* .

Overall, our findings from the interruption logs, survey questions, and interview questions show strong support that the introduction of the FlowLight reduced the cost of interruptions in terms of the overall number as well as their severity.

4.5.2 Increased Awareness of Interruption Cost

Based on the survey responses and interview transcripts, we discovered that after using the FlowLight for some time, participants developed a high degree of awareness for the cost of interruptions:

"It brings more awareness to what people are doing. Sometimes people take it for granted that people are always interruptible. But there is actually a cost or a penalty when you interrupt someone. So, I think just the concept is good because it reminds people that there is sometimes

a good time and a bad time to interrupt people. So, I think just from an awareness campaign, it's valuable as well." (I20)

"The pilot increased the sensitivity to interruption. Team members think more about whether an interrupt is necessary and try to find a suitable time." (S45)

The FlowLight thereby served as a physical reminder for the interruptibility of co-workers in the moment and participants generally respect it and its state:

"It's kind of a like a mood indicator ... so it tells people the state ... of the owner of the light. And then it helps people be more aware or attentive to what my current situation is." (I18)

"I think what really changed is ... a different consciousness about interruptions in our team and also with my colleagues ... I think ... they really respect the light. When it's red I think they don't interrupt." (I11)

Overall, 70% of the 23 interview participants explicitly stated that the FlowLight is respected in their offices and 59.6% of 183 survey respondents agreed that colleagues respected the state of their FlowLight vs. 23.0% that did not (Figure 4.5).

The increased awareness and respect also triggered participants to change their behavior in a variety of ways, ranging from thinking twice before asking, to deferring the interruption, asking before interrupting and changing to a different communication channel, such as email or instant messaging:

"People ask each other if they are available, even when the light is green, even to people with no light. When I see the colleague I want to ask a question ... has a red light, then I wait a while, or write an email." (S77)

"If it's red, I'll send them a message so that when they're no longer busy or something like that, they'll see the message and they can respond to it then ... so it doesn't require an immediate response" (I19)

Fortunately, participants used common sense when working with FlowLights. If a light was red or red blinking participants would still interrupt if the request was urgent:

"Once I go up there [to the person] and I see the light and then I also see that they're pretty intense then I'll push it off unless I really need to get answered to." (I17)

4.5.3 Feeling of Increased Productivity and Self-Motivation

As a further effect of the FlowLight, 58.5% of the survey respondents felt more productive using it, while only 20.1% disagreed with this (Figure 4.5). This feeling of increased productivity often stemmed from the fewer interruptions:

"I definitely think it resulted in less interruptions both in person and via Skype. This resulted in more focus and ability to finish work." (S126)

Another reason for the increased productivity is that the FlowLight serves for some participants as a self-monitoring device that motivates them to become or stay focused, which, however, can also be distracting at times:

"Mostly it has helped as a personal monitor only for me. If I see the light red, I sense I am in the flow and I keep working." (I2)

"When I notice that my light is turning yellow, and I'll feel like, 'Oh yeah, I've been idle' and then I do something ... I think the other way, yeah, there's some effect there too. Like, if I see that it's red, or even flashing red, then I'm like, 'Yeah, I've been very active, or productive, I should keep that going.' At the same time, I think it's also a little bit distracting too. Sometimes just because the light is there, I turn around to check it." (I12)

4.5.4 Costs of Using the FlowLight

While people experienced reduced interruption costs and increase in productivity, there are also costs when starting to use the FlowLight. Especially right after installing it the curiosity of co-workers can lead to an increase of interruptions, which, however, diminishes after a few days:

"People walk by, they see it, they ask me questions, 'What's that? How does it work? What's going on?' like this." (I19)

"Initially there were many people just curious to know what the light is about. This increased the number of interruptions but after few days, people started to respect [it]." (S16)

A few participants also experienced situations in which the FlowLight provoked interruptions, as the green color of the light might be misunderstood as an invitation (observed by 26% of the interviewed participants):

"What I definitely notice is that green is more inviting. So it actually encourages people to come by and say, "Hello" for me at least." (I20)

In some cases, changing the interaction culture might require a mandate from higher up or can even be too expensive:

"The more important issue is for it to work, you have to have people committed to following the light rules, which probably requires engagement of some higher management ... and requires introducing the lights to a wider audience." (I6)

"For us ... the main cost of introducing [it is] that you have to change how you are used to interact with people, that you first have to remember to take a look at the light. That's something that's probably too much for the team. [In] our environment .. it's easier to look at the people than at lights." (I8)

If colleagues choose to ignore the light, especially for unimportant interruptions, it can lead to negative emotions:

"So, for us, what we also heard sometimes is that people have the light red, and others still interrupt them, and they're like, 'Oh no, I have this light red, why did they?' Like it bothers them, and it creates negative emotions almost more than it creates positive emotions..." (I17)

Finally, the public disclosure of the interruptibility status might make people feel exposed at times (8% of survey participants agree, 6% strongly agree) or lead to negative feelings:

"Oh, do other people see that my light is yellow? And are they thinking that I'm not working?" (I12)

Like any new technology, there is a cost to adopting the FlowLight. However, most of the identified costs diminish quickly or can be mitigated by clear direction from management. Overall though participants predominantly stated that the colors of the light were interpreted appropriately and were mostly not concerned about being observed.

4.5.5 Automatic State Changes and Accuracy

The algorithm of the FlowLight caused automatic state changes to indicate a user is, for instance, available for interruptions or busy and not interruptible.

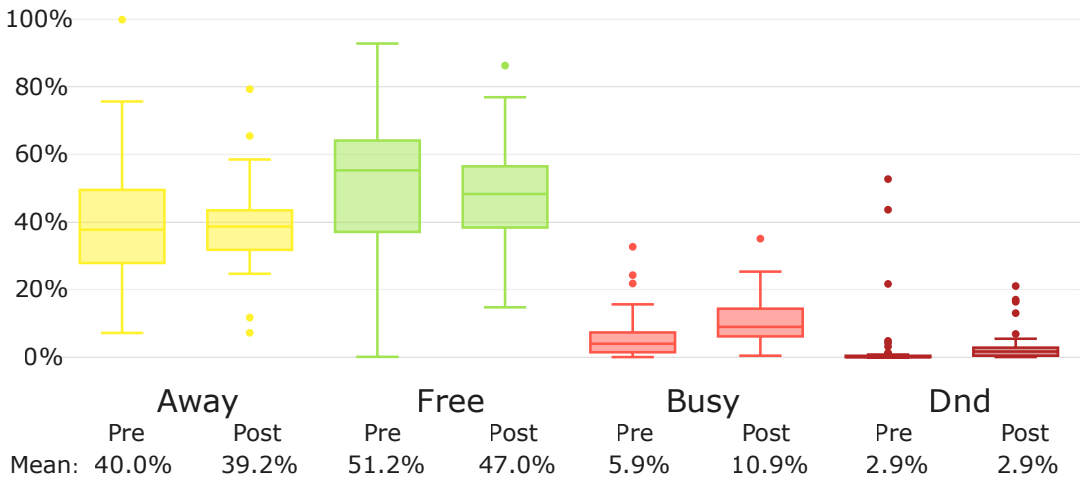


Figure 4.6: Time spent in each state before (pre) and after (post) installation (n=47).

Figure 4.4b illustrates the change in distribution of the number of state changes per participant per day before and after installation. A Wilcoxon signed-rank test with Bonferroni-adjusted alpha levels of .01 per test (.05/5) showed a statistically significant change in the number of state changes (Z =-5.5337, $p \leq .01$) with an increase in state changes from 1.8 before to 8.4 after. This increase shows that the automated algorithm is affecting users' availability status in Skype. Figure 4.6 presents the time spent in each state before and after light installation. Analysis of this data shows a small insignificant decrease in time spent in the available state from 51.2% to 47.0% (Wilcoxon signed-rank test Z =-1.7143, p = .043) and a significant increase in the time spent in the busy state from 5.9% to 10.9% (Z =-3.6403, $p \leq .01$), a very small yet significant difference in do not disturb state (Z =-3.2093, $p \leq .01$), and no significant changes to time spent in the away state. Note that during the before-light period the status was already affected by meetings entered in the calendar, which caused the status to change to busy.

Participants generally agreed that the FlowLight captured their state of availability for interruptions accurately:

"I think it [state representativeness] was actually quite good, because what I found is, if I'm not

working on a critical task, for example, responding to email which usually isn't critically mind provoking. The light would be green and then people would take that opportunity to stop by and see what they needed to talk about. Whereas if I was in the middle of a meeting or if I was more involved in my work, it would turn red and then at that point they might wait for it to turn green. That's my impression." (I20)

Overall, 71.0% of survey respondents agreed that the FlowLight captures their state accurately while only 15.8% disagreed (Figure 4.5). This shows that even an interruptibility measure based on a simple algorithm might be accurate enough to be accepted by users and provide value.

At the same time, interview participants and 64% of our survey respondents mentioned that there are situations where the FlowLight was not representative and accurate, partly stemming from limitations in measuring interruptibility solely with computer interaction data:

"The light was mostly green while debugging code. During debugging, I think interrupts hurt a lot. On the other hand, the light was sometimes red when working on documents / e-mails that do not require too much focus." (S45)

"[The] light captures the movements of the mouse and keyboard, and actually, there are times, which I think of a solution separate from the time, which I implement [it] so ... I'm the most occupied when I think something and usually, I write it on a paper or just keep it on my mind." (I4)

In several cases, participants just changed to setting the state manually when it was not accurate and they wanted to indicate to others that they are available or do not want to be disturbed:

"There was a case when I was reading an article, and I needed a 100% concentration on that, so I just manually changed my status to busy. It was helping me a lot. I think my colleagues are also doing the same when they are engrossed in an article and they want free time, they'll just keep their light busy." (I4)

In fact, 32% of our survey respondents reported to have changed their Skype status (which is linked to the FlowLight) more often after the light was installed, 23% less often and 45% had no changes. With the FlowLight installed, 17% of participants reported to change their status at least once a day, 37% one to

several times a week, and 46% rarely or never. The job role can also affect the accuracy of the FlowLight, especially for managers, administrative assistants, and sales people. For instance, several managers mentioned that interaction was such a core part of their role that they felt they should always be available and turned off the automatic feature.

4.5.6 Continued Usage of FlowLight

Most participants, 82.6% of the 23 interview participants and 79.1% of the 183 survey participants, stated their intention to keep using the FlowLight even after the pilot period. This sentiment is reflected in actual usage data: two months after installing the FlowLight application 85.5% of users remained active (384/449).

Based on online logging of application instances that we started in November 2015, Figure 4.2 shows the number of active FlowLight users per day. The Figure also depicts the start date and relative size for the major pilots (e.g., India #1 started in August '15 and had 80 participants, Norway started in November '14 and had 44).

Note that due to holidays in different locales, vacation, sick days, and travel the number of active users per day is consistently about 70% of the number of unique users over the last month (e.g., a measure of 200 active users per day indicates about 315 number of unique users in the last month).

In spite of most users continuing to use the FlowLight, about 20% of users discontinued usage. There were several reasons that we identified from the interviews and surveys that decreased the benefit of the FlowLight, including the office layout and the visibility of the LED light, the company culture and people ignoring the lights, the initial willingness to use such a system, and the accuracy of the state indicated by the FlowLight. In some cases, the decreased benefit also resulted in participants ceasing to use the FlowLight:

"From my perspective that was something I was against from the first day but as I said I decided to join the pilot because I am a team member. ... From time to time I was looking at it but it

was a little bit discouraging because the color of the light didn't reflect what I was doing and maybe after one week of using it I gave up totally." (I9)

4.5.7 Professional Differences in Using the FlowLight

An analysis of the survey responses with respect to professional roles shows that developers (including testers) and project managers stated more frequently than participants from other working areas that they wanted to continue using the FlowLight , even though not significantly (82% vs 70% on average) and perceived their state to be significantly more accurate (77% vs 60%, t =2.51, p =.01). For project managers, these differences might be explained by the fact that they also reported more often (but not significantly) to manually change their FlowLight status on a daily basis than participants from other work areas (24% vs 16%) and by our experiences gathered during the installation phase, in which managers often asked to disable the automatic mode completely as they wanted to be available for most of their work time. For developers, the differences might be explained by their extensive computer interaction, but future research is needed to confirm this.

4.6 Discussion

The results of our large-scale and long-term study show that the FlowLight can reduce the interruption costs for knowledge workers and can increase the awareness, amongst other benefits. In the following, we discuss implications of our findings, in particular with respect to the combination of the physical indicator with the automatic interruptibility measure, the accuracy of the measure, and the cost of not interrupting. Finally, we discuss threats to validity and limitations of our study.

4.6.1 Reasons for FlowLight's Positive Effects

The FlowLight uses a combination of a physical LED light with an automatic measure based on computer interaction to update the user's interruptibility status.

The findings show that the approach was well adopted and successfully reduced in-person interruption costs. This poses the question if these effects might after all stem solely either from the automatic interruptibility measure or the physical LED light. With respect to the sole use of an automatic interruptibility measure, prior related work that used an automatic measure to update computer-based contact-list style tools, did not find any or the same level of positive effects as our study on both, cost reduction and awareness [Tang et al., 2001, Begole et al., 2004, Lai et al., 2003]. On the other hand, manually maintaining the interruptibility state incurs a high cost as shown by previous research [Milewski and Smith, 2000] and only very few of our users switched to the manual option in cases the algorithm was not accurate enough or they wanted to ensure some undisrupted time. In addition, our findings show that while participants have a high tolerance for the accuracy of the automatic interruptibility status updates, when inaccuracies happen too often, participants also stop using the approach altogether. Overall, this indicates that the combination of the physical LED light and the automatic interruptibility measure is important to provide significant benefits to knowledge workers to use it in the long-term and that it led to the positive impact on awareness and interruption cost found in our study.

4.6.2 Accuracy of Automatic Interruptibility Measure

Participants' high tolerance for the accuracy of the automatic interruptibility measure of the FlowLight poses the question of how accurate the underlying measure has to be to provide sufficient benefit to the user. Over the course of our field study, we adapted the automatic measure two times to account for early user feedback, yet we did not find any significant differences in the effects on interruption cost and behavior. However, we intend to study the relation between accuracy and the effects on interruption cost further in the future.

Also, while participants had a high tolerance, they reported numerous situations in which they observed the status to be set incorrectly. The most frequent situation in which the status is incorrect occurs when participants "think" about something and experience a high cognitive load, yet do not interact with the computer at all. In future work and with the continuously decreasing invasive-

ness of biometric sensors, we plan to extend our approach to integrate biometric sensors, to cover these situations more accurately. We further plan to improve our algorithm by integrating application data, which we were not able to collect in this study due to privacy constraints. Knowing the current application might improve the algorithm's accuracy, e.g. one might be less interruptible while working in a development related program and more while being in an email client. As the nature of work and interactions vary across work areas and job roles, tailoring the algorithm accordingly could further improve its accuracy.

4.6.3 Cost of Not Interrupting

As related work has shown, not all interruptions are bad and some are definitely needed, for instance, to unblock co-workers. By physically indicating knowledge workers as not interruptible (Busy and DnD state), the FlowLight might prevent co-workers from interrupting them for important issues, reducing overall team productivity. The findings of our study on the FlowLight provides evidence that this cost is minimal at best for two reasons. First, a data analysis of the usage logs collected for our study shows that the FlowLight ends up having a significant yet small effect on the time that a knowledge worker is indicated as not interruptible (+5% per day). This indicates that while the FlowLight's automatic algorithm caused a significantly higher number of status updates in Skype compared to manual status updates (an increase from 1.8 to 8.4), these more frequent status updates only minimally changed the knowledge worker's available time by 5%. Second, while the FlowLight increases the awareness of the cost of interruptions, participants still interrupt their co-workers regardless of the FlowLight state if they have an important concern to discuss, as also stated by 35% of our interview participants, without being explicitly asked.

4.6.4 Threats and Limitations

A major threat to the validity of our study is the completeness of the collected data. For instance, we were not able to identify participants across different data sets. While we encouraged participants to share their data and ensured

them that we only use it for research purposes, we could not demand it due to privacy concerns. We were also not able to collect geographic data due to privacy concerns and thus were not able to analyze geographic differences.

Similarly, the accuracy of the interruption logs might be incomplete or not completely accurate. Since interruption logs are based on self-reports, participants might have forgotten to log some interruptions. Also, the work patterns and habits of the days on which they logged interruptions before and after the installation of the FlowLight might have been significantly different, which makes it more difficult to compare the effect of the FlowLight. We tried to mitigate this risk by only including the logs of participants who logged interruptions for more than three days before and three days after and by regularly reminding them to log their interruptions. Furthermore, different participants might have different criteria and judgement standards for logging interruptions. We tried to mitigate this fact by instructing participants to only log external in-person interruptions at work. In addition, by using a paired test that only compares within subject (Wilcoxon signed rank), we mitigate this effect as long as participants did not change their definition of an interruption over time.

We limited the validity threats related to generalizability across individuals and teams by collecting data from 449 participants from twelve countries and with a variety of job roles. As not all participants are native English speakers, there might be a response bias. We tried to mitigate this risk by providing sufficient instructions, opportunity for contacting us if participants had any questions, and also by visiting each major pilot site to introduce and explain the study. Based on the large number and diversity of participants, we observed that responses were not dominantly distributed to extremes, which would indicate that these knowledge workers were particularly biased based on such difficulties. From our in-person experience we can report that with very few exceptions we perceived similar acceptance, respect and in general a very positive perception of the FlowLight across all locations.

Another threat is the influence of the various algorithms on the study results. Since we wanted to ensure that participants are satisfied with the FlowLight and that we take their feedback serious, we evolved the algorithm two times. To

mitigate the risk of a certain bias in the data, we looked for significant differences between populations where we might expect to find them and did not find any.

4.7 Conclusion

In-person interruptions at the workplace can incur a high cost and consume a lot of a knowledge worker's time, if they happen at inopportune moments. While there are several approaches to possibly reduce the interruption costs, little is known about the impact of a physical and automatic interruptibility indicator. In this chapter, we presented FlowLight—an automatic interruptibility indicator in the form of a physical traffic-light like LED—and reported on results from a large-scale and long-term field study with 449 participants from 12 countries. We found that the FlowLight significantly reduced the number of interruptions by 46% (based on 36 interruption logs). We also observed an increased awareness of the potential disruptiveness of interruptions at inopportune moments, which impacts the interaction culture in a positive way, and that our approach can motivate knowledge workers and make them feel more productive (based on 183 survey responses and 23 interview transcripts). We discuss the importance of combining the physical indicator with the automatic interruptibility measure and the high tolerance of participants to the accuracy of the approach. Overall, our study provides deep insights and strong evidence on the very positive effects of the long-term usage of the FlowLight, and the continued usage of the approach by most participants indicates the success of the approach.

4.8 Acknowledgments

The authors would like to thank all study participants. This work was funded in part by SNF.

Bibliography

[Acharya et al., 2006] Acharya, U. R., Joseph, K. P., Kannathal, N., Lim, C. M., and Suri, J. S. (2006). Heart rate variability: a review. *Medical and biological engineering and computing*, 44(12):1031–1051.

[Adamczyk and Bailey, 2004] Adamczyk, P. D. and Bailey, B. P. (2004). If not now, when?: the effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 271–278. ACM.

[Altmann and Trafton, 2002] Altmann, E. M. and Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive science*, 26(1):39–83.

[Altmann and Trafton, 2004] Altmann, E. M. and Trafton, J. G. (2004). Task interruption: Resumption lag and the role of cues. Technical report, MICHIGAN STATE UNIV EAST LANSING DEPT OF PSYCHOLOGY.

[Arroyo and Selker, 2011] Arroyo, E. and Selker, T. (2011). Attention and intention goals can mediate disruption in human-computer interaction. In *IFIP Conference on Human-Computer Interaction*, pages 454–470. Springer.

[Association, 2016] Association, A. H. (2016). Target heart rates.

[Bacchelli and Bird, 2013] Bacchelli, A. and Bird, C. (2013). Expectations, outcomes, and challenges of modern code review. In *Proceedings of the 2013 international conference on software engineering*, pages 712–721. IEEE Press.

[Bahreini et al., 2016] Bahreini, K., Nadolski, R., and Westera, W. (2016). Data fusion for real-time multimodal emotion recognition through webcams and microphones in e-learning. *International Journal of Human–Computer Interaction*, 32(5):415–430.

[Bailey and Iqbal, 2008] Bailey, B. P. and Iqbal, S. T. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 14(4):21.

[Bailey and Konstan, 2006] Bailey, B. P. and Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behavior*, 22(4):685–708.

[Bailey et al., 2001] Bailey, B. P., Konstan, J. A., and Carlis, J. V. (2001). The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *Interact*, volume 1, pages 593–601.

[Beatty, 1982] Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, 91(2):276.

[Begole et al., 2004] Begole, J. B., Matsakis, N. E., and Tang, J. C. (2004). Lilsys: sensing unavailability. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 511–514. ACM.

[Berger, 1929] Berger, H. (1929). Über das elektrenkephalogramm des menschen. *Archiv für psychiatrie und nervenkrankheiten*, 87(1):527–570.

[Beyea, 2007] Beyea, S. C. (2007). Distractions, interruptions, and patient safety. *AORN journal*, 86(1):109–112.

[Bjelica et al., 2011] Bjelica, M. Z., Mrazovac, B., Papp, I., and Teslic, N. (2011). Busy flag just got better: Application of lighting effects in mediating social interruptions. In *MIPRO, 2011 Proceedings of the 34th International Convention*, pages 975–980. IEEE.

[Böhmer et al., 2014] Böhmer, M., Lander, C., Gehring, S., Brumby, D. P., and Krüger, A. (2014). Interrupted by a phone call: exploring designs for lowering the impact of call notifications for smartphone users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3045–3054. ACM.

[Borst et al., 2015] Borst, J. P., Taatgen, N. A., and van Rijn, H. (2015). What makes interruptions disruptive?: A process-model account of the effects of the problem state bottleneck on task interruption and resumption. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 2971–2980. ACM.

[Boucsein, 2012] Boucsein, W. (2012). *Electrodermal activity*. Springer Science & Business Media.

[Brehmer et al., 2012] Brehmer, M., McGrenere, J., Tang, C., and Jacova, C. (2012). Investigating interruptions in the context of computerised cognitive testing for older adults. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2649–2658. ACM.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Butte et al., 2012] Butte, N. F., Ekelund, U., and Westerterp, K. R. (2012). Assessing physical activity using wearable monitors: measures of physical activity. *Medicine & Science in Sports & Exercise*, 44(1S):S5–S12.

[Cades et al., 2007] Cades, D. M., Davis, D. A. B., Trafton, J. G., and Monk, C. A. (2007). Does the difficulty of an interruption affect our ability to resume? In *Proceedings of the human factors and ergonomics society annual meeting*, volume 51, pages 234–238. SAGE Publications Sage CA: Los Angeles, CA.

[Camm et al., 1996] Camm, A. J., Malik, M., Bigger, J., Breithardt, G., Cerutti, S., Cohen, R. J., Coumel, P., Fallen, E. L., Kennedy, H. L., Kleiger, R., et al. (1996). Heart rate variability. standards of measurement, physiological interpretation, and clinical use. *European heart journal*, 17(3):354–381.

[Chen et al., 2007] Chen, D., Hart, J., and Vertegaal, R. (2007). Towards a physiological model of user interruptability. In *IFIP Conference on Human-Computer Interaction*, pages 439–451. Springer.

[Chen and Vertegaal, 2004] Chen, D. and Vertegaal, R. (2004). Using mental load for managing interruptions in physiologically attentive user interfaces. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1513–1516. ACM.

[Chong and Siino, 2006] Chong, J. and Siino, R. (2006). Interruptions on software teams: a comparison of paired and solo programmers. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 29–38. ACM.

[Collet et al., 1997] Collet, C., Vernet-Maury, E., Delhomme, G., and Dittmar, A. (1997). Autonomic nervous system response patterns specificity to basic emotions. *Autonomic Neuroscience: Basic and Clinical*, 62(1):45–57.

[Conard and Marsh, 2010] Conard, M. A. and Marsh, R. (2010). Single and multiple interruptions increase task performance time, but don't affect stress, pressure or flow.

[Czerwinski et al., 2000] Czerwinski, M., Cutrell, E., and Horvitz, E. (2000). Instant messaging: Effects of relevance and timing. In *People and computers XIV: Proceedings of HCI*, volume 2, pages 71–76. British Computer Society.

[Czerwinski et al., 2004] Czerwinski, M., Horvitz, E., and Wilhite, S. (2004). A diary study of task switching and interruptions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 175–182. ACM.

[Electro, 2017] Electro, P. (2017). Equine H7 heart rate sensor belt set. https://www.polar.com/en/products/equine/accessories/equine_H7_heart_rate_sensor_belt_set. [Online; accessed 19-September-2017].

[Embrava, 2016] Embrava (2016). http://www.embrava.com/.

[Fisher and Simmons, 2011] Fisher, R. and Simmons, R. (2011). Smartphone interruptibility using density-weighted uncertainty sampling with reinforcement learning. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, volume 1, pages 436–441. IEEE.

[Fitbit, 2017] Fitbit, I. (2017). Fitbit Charge 2. https://www.fitbit.com/de/charge2. [Online; accessed 19-September-2017].

[Fogarty et al., 2005a] Fogarty, J., Hudson, S. E., Atkeson, C. G., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J. C., and Yang, J. (2005a). Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(1):119–146.

[Fogarty et al., 2005b] Fogarty, J., Ko, A. J., Aung, H. H., Golden, E., Tang, K. P., and Hudson, S. E. (2005b). Examining task engagement in sensor-based statistical models of human interruptibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 331–340. ACM.

[Fogarty et al., 2004] Fogarty, J., Lai, J., and Christensen, J. (2004). Presence versus availability: the design and evaluation of a context-aware communication client. *International Journal of Human-Computer Studies*, 61(3):299–317.

[Fritz et al., 2014] Fritz, T., Begel, A., Müller, S. C., Yigit-Elliott, S., and Züger, M. (2014). Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th international conference on software engineering*, pages 402–413. ACM.

[Gevins et al., 1998] Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., and Rush, G. (1998). Monitoring working memory load during computer-based tasks with eeg pattern recognition methods. *Human factors*, 40(1):79–91.

[Gillie and Broadbent, 1989] Gillie, T. and Broadbent, D. (1989). What makes interruptions disruptive? a study of length, similarity, and complexity. *Psychological research*, 50(4):243–250.

[González and Mark, 2004] González, V. M. and Mark, G. (2004). Constant, constant, multi-tasking craziness: managing multiple working spheres. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 113–120. ACM.

[Goyal and Fussell, 2017] Goyal, N. and Fussell, S. R. (2017). Intelligent interruption management using electro dermal activity based physiological sensor for collaborative sensemaking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):52.

[Grandhi and Jones, 2010] Grandhi, S. and Jones, Q. (2010). Technology-mediated interruption management. *International Journal of Human-Computer Studies*, 68(5):288–306.

[Grandhi and Jones, 2015] Grandhi, S. A. and Jones, Q. (2015). Knock, knock! who' s there? putting the user in control of managing interruptions. *International Journal of Human-Computer Studies*, 79:35–50.

[Grimes et al., 2008] Grimes, D., Tan, D. S., Hudson, S. E., Shenoy, P., and Rao, R. P. (2008). Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 835–844. ACM.

[Guo et al., 2013] Guo, F., Li, Y., Kankanhalli, M. S., and Brown, M. S. (2013). An evaluation of wearable activity monitoring devices. In *Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia*, pages 31–34. ACM.

[Haag et al., 2004] Haag, A., Goronzy, S., Schaich, P., and Williams, J. (2004). Emotion recognition using bio-sensors: First steps towards an automatic system. In *Tutorial and research workshop on affective dialogue systems*, pages 36–48. Springer.

[Haapalainen et al., 2010] Haapalainen, E., Kim, S., Forlizzi, J. F., and Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. In

Proceedings of the 12th ACM international conference on Ubiquitous computing, pages 301–310. ACM.

[Handy, 2005] Handy, T. C. (2005). *Event-related potentials: A methods handbook*. MIT press.

[Hänsel et al., 2018] Hänsel, K., Poguntke, R., Haddadi, H., Alomainy, A., and Schmidt, A. (2018). What to put on the user: Sensing technologies for studies and physiology aware systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 145. ACM.

[Healey and Picard, 2005] Healey, J. A. and Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166.

[Hincapié-Ramos et al., 2011a] Hincapié-Ramos, J. D., Voida, S., and Mark, G. (2011a). A design space analysis of availability-sharing systems. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 85–96. ACM.

[Hincapié-Ramos et al., 2011b] Hincapié-Ramos, J. D., Voida, S., and Mark, G. (2011b). Sharing availability information with interruptme. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 477–478. ACM.

[Hjortskov et al., 2004] Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., and Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology*, 92(1-2):84–89.

[Ho and Intille, 2005] Ho, J. and Intille, S. S. (2005). Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 909–918. ACM.

[Horvitz and Apacible, 2003] Horvitz, E. and Apacible, J. (2003). Learning and reasoning about interruption. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 20–27. ACM.

[Horvitz et al., 2004] Horvitz, E., Koch, P., and Apacible, J. (2004). Busybody: creating and fielding personalized models of the cost of interruption. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 507–510. ACM.

[Horvitz et al., 2002] Horvitz, E., Koch, P., Kadie, C. M., and Jacobs, A. (2002). Coordinate: Probabilistic forecasting of presence and availability. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 224–233. Morgan Kaufmann Publishers Inc.

[Hudson et al., 2003] Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J., and Yang, J. (2003). Predicting human interruptibility with sensors: a wizard of oz feasibility study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 257–264. ACM.

[Iqbal et al., 2004a] Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., and Bailey, B. P. (2004a). Changes in mental workload during task execution. In *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology*.

[Iqbal and Bailey, 2005] Iqbal, S. T. and Bailey, B. P. (2005). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1489–1492. ACM.

[Iqbal and Bailey, 2006] Iqbal, S. T. and Bailey, B. P. (2006). Leveraging characteristics of task structure to predict the cost of interruption. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 741–750. ACM.

[Iqbal and Bailey, 2007] Iqbal, S. T. and Bailey, B. P. (2007). Understanding and developing models for detecting and differentiating breakpoints during

interactive tasks. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 697–706. ACM.

[Iqbal and Bailey, 2008] Iqbal, S. T. and Bailey, B. P. (2008). Effects of intelligent notification management on users and their tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 93–102. ACM.

[Iqbal and Horvitz, 2007] Iqbal, S. T. and Horvitz, E. (2007). Disruption and recovery of computing tasks: field study, analysis, and directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 677–686. ACM.

[Iqbal et al., 2004b] Iqbal, S. T., Zheng, X. S., and Bailey, B. P. (2004b). Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1477–1480. ACM.

[Isaacs et al., 1997] Isaacs, E., Whittaker, S., Frohlich, D., and O'Conaill, B. (1997). Informal communication re-examined: New functions for video in supporting opportunistic encounters. *Video-mediated communication*, 997:459–485.

[Jacob and Karn, 2003] Jacob, R. J. and Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye*, pages 573–605. Elsevier.

[Jones et al., 01] Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python. [Online; accessed 08.01.2018].

[Kapoor and Horvitz, 2007] Kapoor, A. and Horvitz, E. (2007). Principles of lifelong learning for predictive user modeling. *User Modeling 2007*, pages 37–46.

[Kapoor and Horvitz, 2008] Kapoor, A. and Horvitz, E. (2008). Experience sampling for building predictive user models: a comparative study. In *Proceedings*

of the SIGCHI Conference on Human Factors in Computing Systems, pages 657–666. ACM.

[Karvonen and Vuorimaa, 1988] Karvonen, J. and Vuorimaa, T. (1988). Heart rate and exercise intensity during sports activities. *Sports Medicine*, 5(5):303–311.

[Katidioti et al., 2016] Katidioti, I., Borst, J. P., Bierens de Haan, D. J., Pepping, T., van Vugt, M. K., and Taatgen, N. A. (2016). Interrupted by your pupil: An interruption management system based on pupil dilation. *International Journal of Human–Computer Interaction*, 32(10):791–801.

[Kobayashi et al., 2015] Kobayashi, Y., Tanaka, T., Aoki, K., and Fujita, K. (2015). Automatic delivery timing control of incoming email based on user interruptibility. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1779–1784. ACM.

[Komuro et al., 2017] Komuro, K., Fujimoto, Y., and Fujita, K. (2017). Relationship between worker interruptibility and work transitions detected by smartphone. In *International Conference on Human-Computer Interaction*, pages 687–699. Springer.

[Kramer, 1991] Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. *Multiple-task performance*, pages 279–328.

[Kruglanski, 1990] Kruglanski, A. W. (1990). Lay epistemic theory in social-cognitive psychology. *Psychological Inquiry*, 1(3):181–197.

[Lai et al., 2003] Lai, J., Yoshihama, S., Bridgman, T., Podlaseck, M., Chou, P. B., and Wong, D. C. (2003). Myteam: Availability awareness through the use of sensor data. In *INTERACT*.

[Lee and Tan, 2006] Lee, J. C. and Tan, D. S. (2006). Using a low-cost electroencephalograph for task classification in hci research. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, pages 81–90. ACM.

[Lemaire, 1996] Lemaire, P. (1996). The role of working memory resources in simple cognitive arithmetic. *European Journal of Cognitive Psychology*, 8(1):73–104.

[Liaw et al., 2002] Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

[Maaoui et al., 2010] Maaoui, C., Pruski, A., and Abdat, F. (2010). *Emotion recognition through physiological signals for human-machine communication*. INTECH Open Access Publisher.

[Manoilov, 2007] Manoilov, P. (2007). Eye-blinking artefacts analysis. In *Proceedings of the 2007 international conference on Computer systems and technologies*, page 52. ACM.

[Mark et al., 2016a] Mark, G., Czerwinski, M., Iqbal, S., and Johns, P. (2016a). Workplace indicators of mood: Behavioral and cognitive correlates of mood among information workers. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 29–36. ACM.

[Mark et al., 2005] Mark, G., Gonzalez, V. M., and Harris, J. (2005). No task left behind?: examining the nature of fragmented work. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 321–330. ACM.

[Mark et al., 2008] Mark, G., Gudith, D., and Klocke, U. (2008). The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 107–110. ACM.

[Mark et al., 2014] Mark, G., Iqbal, S. T., Czerwinski, M., and Johns, P. (2014). Bored mondays and focused afternoons: the rhythm of attention and online activity in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3025–3034. ACM.

[Mark et al., 2016b] Mark, G., Iqbal, S. T., Czerwinski, M., Johns, P., Sano, A., and Lutchyn, Y. (2016b). Email duration, batching and self-interruption:

Patterns of email use on productivity and stress. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1717–1728. ACM.

[Mathan et al., 2007] Mathan, S., Whitlow, S., Dorneich, M., Ververs, P., and Davis, G. (2007). Neurophysiological estimation of interruptibility: Demonstrating feasibility in a field context. In *In Proceedings of the 4th International Conference of the Augmented Cognition Society*, pages 51–58.

[McCarney et al., 2007] McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., and Fisher, P. (2007). The hawthorne effect: a randomised, controlled trial. *BMC medical research methodology*, 7(1):30.

[McFarlane, 2002] McFarlane, D. (2002). Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction*, 17(1):63–139.

[Melillo et al., 2011] Melillo, P., Bracale, M., and Pecchia, L. (2011). Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due to university examination. *Biomedical engineering online*, 10(1):96.

[Meyer et al., 2017a] Meyer, A. N., Barton, L. E., Murphy, G. C., Zimmermann, T., and Fritz, T. (2017a). The work life of developers: Activities, switches and perceived productivity. *IEEE Transactions on Software Engineering*.

[Meyer et al., 2014] Meyer, A. N., Fritz, T., Murphy, G. C., and Zimmermann, T. (2014). Software developers' perceptions of productivity. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*, pages 19–29. ACM.

[Meyer et al., 2017b] Meyer, A. N., Murphy, G. C., Zimmermann, T., and Fritz, T. (2017b). Retrospecting on work and productivity: A study on self-monitoring software developers' work. *Proc. ACM Hum.-Comput. Interact.*, (CSCW):79:1–79:24.

[Microsoft, 2017] Microsoft (2017). Microsoft Graph. https://graph.microsoft.io. [Online; accessed 19-September-2017].

[Milewski and Smith, 2000] Milewski, A. E. and Smith, T. M. (2000). Providing presence cues to telephone users. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 89–96. ACM.

[Mirza et al., 2011] Mirza, H. T., Chen, L., Chen, G., Hussain, I., and He, X. (2011). Switch detector: an activity spotting system for desktop. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2285–2288. ACM.

[Monk et al., 2008] Monk, C. A., Trafton, J. G., and Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied*, 14(4):299.

[Montgomery-Downs et al., 2012] Montgomery-Downs, H. E., Insana, S. P., and Bond, J. A. (2012). Movement toward a novel activity monitoring device. *Sleep and Breathing*, 16(3):913–917.

[Mulder, 1992] Mulder, L. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological psychology*, 34(2):205–236.

[Müller and Fritz, 2015] Müller, S. C. and Fritz, T. (2015). Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress. In *Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on*, volume 1, pages 688–699. IEEE.

[Nair et al., 2005] Nair, R., Voida, S., and Mynatt, E. D. (2005). Frequency-based detection of task switches. In *Proceedings of the 19th British HCI Group Annual Conference*, volume 2, pages 94–99.

[Nardi et al., 2000] Nardi, B. A., Whittaker, S., and Bradner, E. (2000). Interaction and outeraction: instant messaging in action. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 79–88. ACM.

[Nourbakhsh et al., 2012] Nourbakhsh, N., Wang, Y., Chen, F., and Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, pages 420–423. ACM.

[Parnin and Rugaber, 2011] Parnin, C. and Rugaber, S. (2011). Resumption strategies for interrupted programming tasks. *Software Quality Journal*, 19(1):5–34.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Peper et al., 2007] Peper, E., Harvey, R., Lin, I.-M., Tylova, H., and Moss, D. (2007). Is there more to blood volume pulse than heart rate variability, respiratory sinus arrhythmia, and cardiorespiratory synchrony? *Biofeedback*, 35(2).

[Pilcher et al., 1997] Pilcher, J. J., Ginter, D. R., and Sadowsky, B. (1997). Sleep quality versus sleep quantity: relationships between sleep and measures of health, well-being and sleepiness in college students. *Journal of psychosomatic research*, 42(6):583–596.

[Richter et al., 1998] Richter, P., Wagner, T., Heger, R., and Weise, G. (1998). Psychophysiological analysis of mental load during driving on rural roads-a quasi-experimental field study. *Ergonomics*, 41(5):593–609.

[Rokach and Maimon, 2005] Rokach, L. and Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4):476–487.

[Rosekind et al., 2010] Rosekind, M. R., Gregory, K. B., Mallis, M. M., Brandt, S. L., Seal, B., and Lerner, D. (2010). The cost of poor sleep: workplace pro-

ductivity loss and associated costs. *Journal of Occupational and Environmental Medicine*, 52(1):91–98.

[Rosenberger et al., 2016] Rosenberger, M. E., Buman, M. P., Haskell, W. L., McConnell, M. V., and Carstensen, L. L. (2016). 24 hours of sleep, sedentary behavior, and physical activity with nine wearable devices. *Medicine and science in sports and exercise*, 48(3):457.

[Rule et al., 2015] Rule, A., Tabard, A., Boyd, K., and Hollan, J. (2015). Restoring the context of interrupted work with desktop thumbnails. In *37th Annual Meeting of the Cognitive Science Society*.

[Samara et al., 2017] Samara, A., Galway, L., Bond, R., and Wang, H. (2017). Tracking and evaluation of pupil dilation via facial point marker analysis. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, pages 2037–2043. IEEE.

[Sammut and Webb, 2011] Sammut, C. and Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

[Sano and Picard, 2013] Sano, A. and Picard, R. W. (2013). Stress recognition using wearable sensors and mobile phones. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 671–676. IEEE.

[Sarkar and Parnin, 2017] Sarkar, S. and Parnin, C. (2017). Characterizing and predicting mental fatigue during programming tasks. In *Proceedings of the 2nd International Workshop on Emotion Awareness in Software Engineering*, pages 32–37. IEEE Press.

[Schmidt and Walach, 2000] Schmidt, S. and Walach, H. (2000). Electrodermal activity (eda)-state-of-the-art measurement and techniques for parapsychological purposes. *the Journal of Parapsychology*, 64(2):139.

[Shrot et al., 2014] Shrot, T., Rosenfeld, A., Golbeck, J., and Kraus, S. (2014). Crisp: an interruption management algorithm based on collaborative filtering.

In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3035–3044. ACM.

[Snyder et al., 2015] Snyder, J., Matthews, M., Chien, J., Chang, P. F., Sun, E., Abdullah, S., and Gay, G. (2015). Moodlight: Exploring personal and social implications of ambient display of biosensor data. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 143–153. ACM.

[Stern et al., 2011] Stern, H., Pammer, V., and Lindstaedt, S. N. (2011). A preliminary study on interruptibility detection based on location and calendar information. *Proc. CoSDEO*, 11.

[Storey et al., 2017] Storey, M.-A., Zagalsky, A., Figueira Filho, F., Singer, L., and German, D. M. (2017). How social and communication channels shape and challenge a participatory culture in software development. *IEEE Transactions on Software Engineering*, 43(2):185–204.

[Sweller, 1994] Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4):295–312.

[Sweller, 2011] Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier.

[Sykes, 2011] Sykes, E. R. (2011). Interruptions in the workplace: A case study to reduce their effects. *International Journal of Information Management*, 31(4):385–394.

[Tanaka and Fujita, 2011] Tanaka, T. and Fujita, K. (2011). Study of user interruptibility estimation based on focused application switching. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 721–724. ACM.

[Tang et al., 2001] Tang, J. C., Yankelovich, N., Begole, J., Van Kleek, M., Li, F., and Bhalodia, J. (2001). Connexus to awarenex: extending awareness to

mobile users. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–228. ACM.

[Tani and Yamada, 2013] Tani, T. and Yamada, S. (2013). Estimating user interruptibility by measuring table-top pressure. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 1707–1712. ACM.

[Trafton et al., 2003] Trafton, J. G., Altmann, E. M., Brock, D. P., and Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies*, 58(5):583–603.

[Turner et al., 2015] Turner, L. D., Allen, S. M., and Whitaker, R. M. (2015). Interruptibility prediction for ubiquitous systems: conventions and new directions from a growing field. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 801–812. ACM.

[van Solingen et al., 1998] van Solingen, R., Berghout, E., and van Latum, F. (1998). Interrupts: just a minute never is. *IEEE software*, 15(5):97–103.

[Vasilescu et al., 2016] Vasilescu, B., Blincoe, K., Xuan, Q., Casalnuovo, C., Damian, D., Devanbu, P., and Filkov, V. (2016). The sky is not the limit: multitasking across github projects. In *Proceedings of the 38th International Conference on Software Engineering*, pages 994–1005. ACM.

[Veltman and Gaillard, 1998] Veltman, J. and Gaillard, A. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, 41(5):656–669.

[Vidaček et al., 1986] Vidaček, S., Kaliterna, L., Radošević-Vidaček, B., and Folkard, S. (1986). Productivity on a weekly rotating shift system: circadian adjustment and sleep deprivation effects? *Ergonomics*, 29(12):1583–1590.

[Vinkers et al., 2013] Vinkers, C. H., Penning, R., Hellhammer, J., Verster, J. C., Klaessens, J. H., Olivier, B., and Kalkman, C. J. (2013). The effect of stress on core and peripheral body temperature in humans. *Stress*, 16(5):520–530.

[Visuri et al., 2017] Visuri, A., van Berkel, N., Luo, C., Goncalves, J., Ferreira, D., and Kostakos, V. (2017). Predicting interruptibility for manual data collection: A cluster-based user model.

[Vorburger et al., 2011] Vorburger, P., Bernstein, A., and Zurfluh, A. (2011). Interruptability prediction using motion detection. In *1st Int. Workshop on Managing Context Information in Mobile and Pervasive Environments MCMP-05*.

[Wang et al., 2017] Wang, R., Blackburn, G., Desai, M., Phelan, D., Gillinov, L., Houghtaling, P., and Gillinov, M. (2017). Accuracy of wrist-worn heart rate monitors. *Jama cardiology*, 2(1):104–106.

[Wijsman et al., 2011] Wijsman, J., Grundlehner, B., Liu, H., Hermens, H., and Penders, J. (2011). Towards mental stress detection using wearable physiological sensors. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 1798–1801. IEEE.

[Wilson, 2002] Wilson, G. F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, 12(1):3–18.

[Witten et al., 2016] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[Xhyheri et al., 2012] Xhyheri, B., Manfrini, O., Mazzolini, M., Pizzi, C., and Bugiardini, R. (2012). Heart rate variability today. *Progress in cardiovascular diseases*, 55(3):321–331.

[Zeigarnik, 1927] Zeigarnik, B. G. (1927). *Das Behalten erledigter und unerledigter Handlungen, Inaugural-Dissertation... von Bluma Zeigarnik,...* J. Springer.

[Züger et al., 2017] Züger, M., Corley, C., Meyer, A. N., Li, B., Fritz, T., Shepherd, D., Augustine, V., Francis, P., Kraft, N., and Snipes, W. (2017). Reducing

interruptions at work: A large-scale field study of flowlight. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 61–72. ACM.

[Züger and Fritz, 2015] Züger, M. and Fritz, T. (2015). Interruptibility of software developers and its prediction using psycho-physiological sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2981–2990. ACM.

Curriculum Vitae

Personal Information

Name	Manuela Züger
Nationality	Swiss
Date of Birth	April 21, 1988
Place of Birth	Frauenfeld, Switzerland

Education

2014 – 2018	PhD studies in Informatics Department of Informatics University of Zurich, Switzerland
2011 – 2013	Master of Science in Informatics Department of Informatics University of Zurich, Switzerland
2007 – 2011	Bachelor of Science in Informatics Department of Informatics University of Zurich, Switzerland